Red Hat Ceph Storage 7

配置指南

Red Hat Ceph Storage 的配置设置

Last Updated: 2024-06-14

Red Hat Ceph Storage 7 配置指南

Red Hat Ceph Storage 的配置设置

法律通告

Copyright © 2024 Red Hat, Inc.

The text of and illustrations in this document are licensed by Red Hat under a Creative Commons Attribution–Share Alike 3.0 Unported license ("CC-BY-SA"). An explanation of CC-BY-SA is available at
http://creativecommons.org/licenses/by-sa/3.0/
. In accordance with CC-BY-SA, if you distribute this document or an adaptation of it, you must provide the URL for the original version.

Red Hat, as the licensor of this document, waives the right to enforce, and agrees not to assert, Section 4d of CC-BY-SA to the fullest extent permitted by applicable law.

Red Hat, Red Hat Enterprise Linux, the Shadowman logo, the Red Hat logo, JBoss, OpenShift, Fedora, the Infinity logo, and RHCE are trademarks of Red Hat, Inc., registered in the United States and other countries.

Linux ® is the registered trademark of Linus Torvalds in the United States and other countries.

Java ® is a registered trademark of Oracle and/or its affiliates.

XFS ® is a trademark of Silicon Graphics International Corp. or its subsidiaries in the United States and/or other countries.

MySQL ® is a registered trademark of MySQL AB in the United States, the European Union and other countries.

Node.js ® is an official trademark of Joyent. Red Hat is not formally related to or endorsed by the official Joyent Node.js open source or commercial project.

The OpenStack ® Word Mark and OpenStack logo are either registered trademarks/service marks or trademarks/service marks of the OpenStack Foundation, in the United States and other countries and are used with the OpenStack Foundation's permission. We are not affiliated with, endorsed or sponsored by the OpenStack Foundation, or the OpenStack community.

All other trademarks are the property of their respective owners.

摘要

本文档提供有关在引导时配置 Red Hat Ceph Storage 和运行时的信息。它还提供配置参考信息。 红帽致力于替换我们的代码、文档和 Web 属性中存在问题的语言。我们从这四个术语开始：master、slave、黑名单和白名单。由于此项工作十分艰巨，这些更改将在即将推出的几个发行版本中逐步实施。详情请查看 CTO Chris Wright 信息。

目录

第 1 章 CEPH 配置的基础知识

作为存储管理员，您需要基本了解如何查看 Ceph 配置以及如何为 Red Hat Ceph Storage 集群设置 Ceph 配置选项。您可以在运行时查看和设置 Ceph 配置选项。

先决条件

- 安装 Red Hat Ceph Storage 软件。

1.1. CEPH 配置

所有 Red Hat Ceph Storage 集群都有一个配置，它定义：

- 集群身份
- 身份验证设置
- Ceph 守护进程
- 网络配置
- 节点名称和地址
- keyring 的路径
- 到 OSD 日志文件的路径
- 其他运行时选项

部署工具（如 **cephadm** ）通常会为您创建初始 Ceph 配置文件。但是，如果您想在没有使用部署工具的情况下引导 Red Hat Ceph Storage 集群，您可以自己创建一个。

其它资源

- 有关 **cephadm** 和 Ceph 编配器的更多信息，请参阅 *Red Hat Ceph Storage Operations 指南* 。

1.2. CEPH 配置数据库

Ceph Monitor 管理 Ceph 选项的配置数据库，通过存储整个存储集群的配置选项来集中管理配置管理。通过将 Ceph 配置集中到一个数据库中，从而简化了存储集群管理。

Ceph 用于设置选项的优先级顺序是：

- 编译的默认值
- Ceph 集群配置数据库
- 本地 **ceph.conf** 文件
- 运行时覆盖，使用 **ceph daemon *DAEMON-NAME* config set** 或 **ceph tell *DAEMON-NAME* injectargs** 命令

仍然可在本地 Ceph 配置文件中定义一些 Ceph 选项，默认为 **/etc/ceph/ceph.conf**。但是，对于 Red Hat Ceph Storage 7，**ceph.conf** 已被弃用。

cephadm 使用基本的 **ceph.conf** 文件，该文件仅包含连接到 Ceph 监控器、身份验证和获取配置信息的最小选项集合。在大多数情况下，**cephadm** 仅使用 **mon_host** 选项。要避免仅将 **ceph.conf** 用于 **mon_host** 选项，请使用 DNS SRV 记录来通过 monitor 执行操作。

重要

红帽建议您使用 **assimilate-conf** 管理命令将有效的选项从 **ceph.conf** 文件移至配置数据库中。有关 **similate-conf** 的更多信息，请参阅管理命令。

Ceph 允许您在运行时更改守护进程的配置。通过启用或禁用 debug 设置，此功能可用于增加或减少日志输出，甚至可用于运行时优化。

注意

当配置数据库和 Ceph 配置文件中存在相同的选项时，配置数据库选项的优先级低于 Ceph 配置文件中所设置的内容。

部分和掩码

正如您可以在全局范围内、或针对每个守护进程或针对 Ceph 配置文件内的特点守护进程配置 Ceph 选项一样，您也可以根据以下部分在配置数据库中配置 Ceph 选项。

部分	描述
global	对所有守护进程和客户端有效。
mon	影响所有 Ceph 监控器。
mgr	影响所有 Ceph 管理器。
osd	影响所有 Ceph OSD。
mds	影响所有 Ceph 元数据服务器。
client	影响所有 Ceph 客户端，包括挂载的文件系统、块设备和 RADOS 网关。

Ceph 配置选项可以带有关联的掩码。这些掩码可以进一步限制选项应用到的守护进程或客户端。

掩码有两种形式：

type:location

type 是 CRUSH 属性，如 **rack** 或 **host**。**location** 是属性类型的值。例如，**host:foo** 将选项限制为在 **foo** 主机上运行的守护进程或客户端。

示例

```
ceph config set osd/host:magna045 debug_osd 20
```

class:device-class

device-class 是 CRUSH 设备类的名称，如 **hdd** 或 **ssd**。例如，**class:ssd** 会将选项限制为仅限由固态硬盘 (SSD) 支持的 Ceph OSD。这个掩码对客户端的非 OSD 守护进程没有影响。

示例

```
ceph config set osd/class:hdd osd_max_backfills 8
```

管理命令

可以通过子命令 **ceph config *ACTION*** 管理 Ceph 配置数据库。以下是您可以执行的操作：

ls

列出可用的配置选项。

dump

转储存储集群选项的整个配置数据库。

get *WHO*

转储特定守护进程或客户端的配置。例如，*WHO* 可以是一个守护进程，如 **mds.a**。

set *WHO OPTION VALUE*

在 Ceph 配置数据库中设置配置选项，其中 *WHO* 是目标守护进程，*OPTION* 是要设置的选项，*VALUE* 是所需的值。

show *WHO*

为一个正在运行的守护进程显示报告的运行配置。如果使用了一个本地的配置文件，或选项已被命令行或在运行时被覆盖，则这些选项可能与 Ceph Monitor 存储的不同。另外，选项值的来源会报告为输出的一部分。

assimilate-conf -i *INPUT_FILE* -o *OUTPUT_FILE*

从 *INPUT_FILE* 中模拟配置文件，并将任何有效的选项移到 Ceph monitor 的配置数据库中。任何无法被识别、无效或无法由 Ceph Monitor 返回的选项在 *OUTPUT_FILE* 中存储的缩写配置文件中控制。此命令对于从旧配置文件迁移到集中式配置数据库非常有用。请注意，如果您使用一个配置，而监控器或其他守护进程对于同一组选项设置了不同的配置值，则最终结果将取决于文件使用的顺序。

help *OPTION* -f json-pretty

使用 JSON 格式输出显示特定 *OPTION* 的帮助信息。

其它资源

- 有关命令的更多信息，请参阅*在运行时设置特定的配置*。

1.3. 使用 CEPH 元变量

Metavariables 简化了 Ceph 存储集群配置。在配置值中设置 metavariable 时，Ceph 会将 metavariable 扩展为 Concrete 值。

在 Ceph 配置文件的 **[global]**, **[osd]**, **[mon]**, 或 **[client]** 部分中使用时，Mtavariables 非常强大。但是，您也可以在管理套接字中使用它们。Ceph 元变量与 Bash shell 扩展类似。

Ceph 支持以下元变量：

$cluster

描述

扩展至 Ceph 存储集群名称。在同一硬件上运行多个 Ceph 存储群集时很有用。

示例

/etc/ceph/$cluster.keyring

默认

ceph

$type

描述

根据即时守护进程的类型，扩展至 **osd** 或 **mon** 之一。

示例

/var/lib/ceph/$type

$id

描述

扩展至守护进程标识符。对于 **osd.0**，这将是 **0**。

示例

/var/lib/ceph/$type/$cluster-$id

$host

描述

扩展至即时守护进程的主机名。

$name

描述

扩展至 **$type.$id**。

示例

/var/run/ceph/$cluster-$name.asok

1.4. 在运行时查看 CEPH 配置

可以在启动时或运行时查看 Ceph 配置文件。

先决条件

- Ceph 节点的根级别访问权限。
- 对管理密钥环的访问。

流程

1. 要查看运行时配置，请登录到运行守护进程的 Ceph 节点，并执行：

   语法

   ```
   ceph daemon DAEMON_TYPE.ID config show
   ```

   要查看 **osd.0** 的配置，您可以登录到包含 **osd.0** 的节点并执行以下命令：

示例

```
[root@osd ~]# ceph daemon osd.0 config show
```

2. 如需附加选项，指定守护进程和**帮助**。

示例

```
[root@osd ~]# ceph daemon osd.0 help
```

1.5. 在运行时查看特定配置

Red Hat Ceph Storage 的配置设置可在 Ceph 监控节点运行时查看。

先决条件

- 一个正在运行的 Red Hat Ceph Storage 集群。
- Ceph 监控节点的根级别访问权限.

流程

1. 登录 Ceph 节点并执行：

语法

```
ceph daemon DAEMON_TYPE.ID config get PARAMETER
```

示例

```
[root@mon ~]# ceph daemon osd.0 config get public_addr
```

1.6. 在运行时设置特定的配置

要在运行时设置特定的 Ceph 配置，请使用 **ceph config set** 命令。

先决条件

- 一个正在运行的 Red Hat Ceph Storage 集群。
- 对 Ceph monitor 或 OSD 节点的 root 级别访问权限。

流程

1. 在所有 monitor 或 OSD 守护进程上设置配置：

语法

```
ceph config set DAEMON CONFIG-OPTION VALUE
```

示例

```
[root@mon ~]# ceph config set osd debug_osd 10
```

2. 验证是否设置了选项和值：

示例

```
[root@mon ~]# ceph config dump
osd      advanced debug_osd  10/10
```

- 在所有守护进程中删除配置选项：

  语法

  ```
  ceph config rm DAEMON CONFIG-OPTION VALUE
  ```

  示例

  ```
  [root@mon ~]# ceph config rm osd debug_osd
  ```

- 为特定守护进程设置配置：

  语法

  ```
  ceph config set DAEMON.DAEMON-NUMBER CONFIG-OPTION VALUE
  ```

  示例

  ```
  [root@mon ~]# ceph config set osd.0 debug_osd 10
  ```

- 验证是否为指定守护进程设置配置：

  示例

  ```
  [root@mon ~]# ceph config dump
  osd.0      advanced debug_osd     10/10
  ```

- 删除特定守护进程的配置：

  语法

  ```
  ceph config rm DAEMON.DAEMON-NUMBER CONFIG-OPTION
  ```

  示例

  ```
  [root@mon ~]# ceph config rm osd.0 debug_osd
  ```

注意

如果您使用不支持从配置数据库读取选项的客户端，或者您仍需要使用 **ceph.conf** 更改集群配置，请运行以下命令：

```
ceph config set mgr mgr/cephadm/manage_etc_ceph_ceph_conf false
```

您必须在存储集群中维护并分发 **ceph.conf** 文件。

1.7. OSD 内存目标

BlueStore 将 OSD 堆内存使用量保留在指定目标大小下，并使用 **osd_memory_target** 配置选项。

选项 **osd_memory_target** 根据系统中可用的 RAM 来设置 OSD 内存。当 TCMalloc 配置为内存分配器，BlueStore 中的 **bluestore_cache_autotune** 选项设为 **true** 时，则使用此选项。

当块设备速度较慢时（例如，传统的硬盘驱动器），Ceph OSD 内存缓存更为重要，因为缓存命中的好处要高于固态硬盘的情况。但是，这需要考虑 OSD 与其他服务共处的情况，比如在超融合基础架构 (HCI) 或其他应用程序中。

1.7.1. 设置 OSD 内存目标

使用 **osd_memory_target** 选项，为存储集群中的所有 OSD 或特定 OSD 设置最大内存阈值。带有 **osd_memory_target** 选项被设置为 16 的 OSD 最多可使用 16 GB 内存。

注意

单个 OSD 的配置选项会优先于对所有 OSD 的设置。

先决条件

- 一个正在运行的 Red Hat Ceph Storage 集群。
- 对存储集群中所有主机的根级别访问权限。

流程

- 为存储集群中的所有 OSD 设置 **osd_memory_target** ：

  语法

  ```
  ceph config set osd osd_memory_target VALUE
  ```

  VALUE 是要分配给存储集群中每个 OSD 的内存数量。

- 为存储集群中的特定 OSD 设置 **osd_memory_target** ：

  语法

  ```
  ceph config set osd.id osd_memory_target VALUE
  ```

  .id 是 OSD 的 ID， *VALUE* 是要分配给指定 OSD 的内存数量。例如，将 ID 为 8 的 OSD 配置为使用最多 16 GBytes 内存：

示例

```
[ceph: root@host01 /]# ceph config set osd.8 osd_memory_target 16G
```

- 要设置单个 OSD 使用一个最大内存量，并将其余 OSD 配置为使用其他数量，请首先指定单个 OSD：

示例

```
[ceph: root@host01 /]# ceph config set osd osd_memory_target 16G
[ceph: root@host01 /]# ceph config set osd.8 osd_memory_target 8G
```

其他资源

- 要配置 Red Hat Ceph Storage 以自动调优 OSD 内存用量，请参阅 *操作指南*中的*自动调整 OSD 内存*。

1.8. 自动调优 OSD 内存

OSD 守护进程根据 **osd_memory_target** 配置选项调整内存消耗。选项 **osd_memory_target** 根据系统中可用的 RAM 来设置 OSD 内存。

如果 Red Hat Ceph Storage 部署在不与其他服务共享内存的专用节点上，**cephadm** 会自动根据 RAM 总量和部署的 OSD 数量自动调整每个 OSD 消耗。

重要

默认情况下，Red Hat Ceph Storage 集群中的 **osd_memory_target_autotune** 参数设置为 **true**。

语法

```
ceph config set osd osd_memory_target_autotune true
```

Cephadm 以一个 **mgr/cephadm/autotune_memory_target_ratio** 分数开头，默认为系统总 RAM 的 **0.7**，这会减小非自动tuned守护进程（如 non-OSDS）以及 **osd_memory_target_autotune** 为 false 的 OSD，然后划分剩余的 OSD。

osd_memory_target 参数计算如下：

语法

```
osd_memory_target = TOTAL_RAM_OF_THE_OSD * (1048576) * (autotune_memory_target_ratio) / NUMBER_OF_OSDS_IN_THE_OSD_NODE - (SPACE_ALLOCATED_FOR_OTHER_DAEMONS)
```

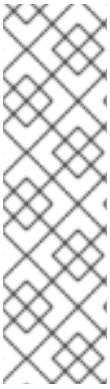
SPACE_ALLOCATED_FOR_OTHER_DAEMONS 可能包括以下守护进程空间分配：

- Alertmanager: 1 GB
- Grafana: 1 GB
- Ceph Manager：4 GB

- Ceph Monitor: 2 GB
- Node-exporter: 1 GB
- Prometheus: 1 GB

例如，如果节点有 24 个 OSD 且具有 251 GB RAM 空间，则 **osd_memory_target** 为 **7860684936**。

最终目标反映在带有选项的配置数据库中。您可以从 **ceph orch ps** 输出的 **MEM LIMIT** 列下查看各个守护进程使用的限值和当前内存。



注意

osd_memory_target_autotune true 的默认设置不适用于计算和 Ceph 存储服务在一起的超融合基础架构。在超融合基础架构中，**autotune_memory_target_ratio** 可以设置为 **0.2**，以减少 Ceph 的内存消耗。

示例

```
[ceph: root@host01 /]# ceph config set mgr
mgr/cephadm/autotune_memory_target_ratio 0.2
```

您可以为存储集群中的 OSD 手动设置特定内存目标。

示例

```
[ceph: root@host01 /]# ceph config set osd.123 osd_memory_target 7860684936
```

您可以为存储集群中的 OSD 主机手动设置特定内存目标。

语法

```
ceph config set osd/host:HOSTNAME osd_memory_target TARGET_BYTES
```

示例

```
[ceph: root@host01 /]# ceph config set osd/host:host01 osd_memory_target 1000000000
```

注意

启用 **osd_memory_target_autotune** 覆盖现有的手动 OSD 内存目标设置。要防止守护进程内存被调整（即使启用了 **osd_memory_target_autotune** 选项或启用了其他类似的选项），在主机上设置 **_no_autotune_memory** 标签。

语法

```
ceph orch host label add HOSTNAME _no_autotune_memory
```

您可以通过禁用 autotune 选项并设置特定内存目标，从内存自动调整 OSD 中排除。

示例

-

```
[ceph: root@host01 /]# ceph config set osd.123 osd_memory_target_autotune false
[ceph: root@host01 /]# ceph config set osd.123 osd_memory_target 16G
```

1.9. MDS 内存缓存限制

MDS 服务器将其元数据保留在一个单独的存储池中（名为 **cephfs_metadata**），并且是 Ceph OSD 的用户。对于 Ceph 文件系统，MDS 服务器必须支持整个 Red Hat Ceph Storage 集群，而不支持存储集群中的单个存储设备，因此它们的内存要求可能会非常显著，特别是当工作负载包含小时时，数据元数据的比例更大。

例如，将 **mds_cache_memory_limit** 设置为 2000000000

```
ceph_conf_overrides:
  mds:
    mds_cache_memory_limit=2000000000
```

注意

对于具有元数据密集型工作负载的大型 Red Hat Ceph Storage 集群，请不要将 MDS 服务器与其他内存密集型服务位于同一个节点上，这样做可让您将更多内存分配给 MDS，例如，分配大于 100 GB 的内存。

其它资源

- 请参阅 *Red Hat Ceph Storage 文件系统指南* 中的*存储服务器缓存大小限制*。
- 有关特定选项描述和使用，请参阅配置 *选项中的* 常规 Ceph 配置选项。

第 2 章 CEPH 网络配置

作为存储管理员，您必须了解 Red Hat Ceph Storage 集群要在其中运行的网络环境，并相应地配置 Red Hat Ceph Storage。了解并配置 Ceph 网络选项可以确保整个存储集群的最佳性能和可靠性。

先决条件

- 网络连接。
- 安装 Red Hat Ceph Storage 软件。

2.1. CEPH 的网络配置

网络配置对于构建高性能 Red Hat Ceph Storage 集群至关重要。Ceph 存储集群不代表 Ceph 客户端执行请求路由或分配请求。相反，Ceph 客户端直接向 Ceph OSD 守护进程发出请求。Ceph OSD 代表 Ceph 客户端执行数据复制，这意味着复制和其他因素对 Ceph 存储集群的网络造成额外的负载。

Ceph 有一个网络配置要求适用于所有守护进程。Ceph 配置文件必须为每个守护进程指定 **host**。

一些部署实用程序（如 **cephadm** ）会为您创建一个配置文件。如果部署实用程序为您设置这些值，则不要设置这些值。

重要

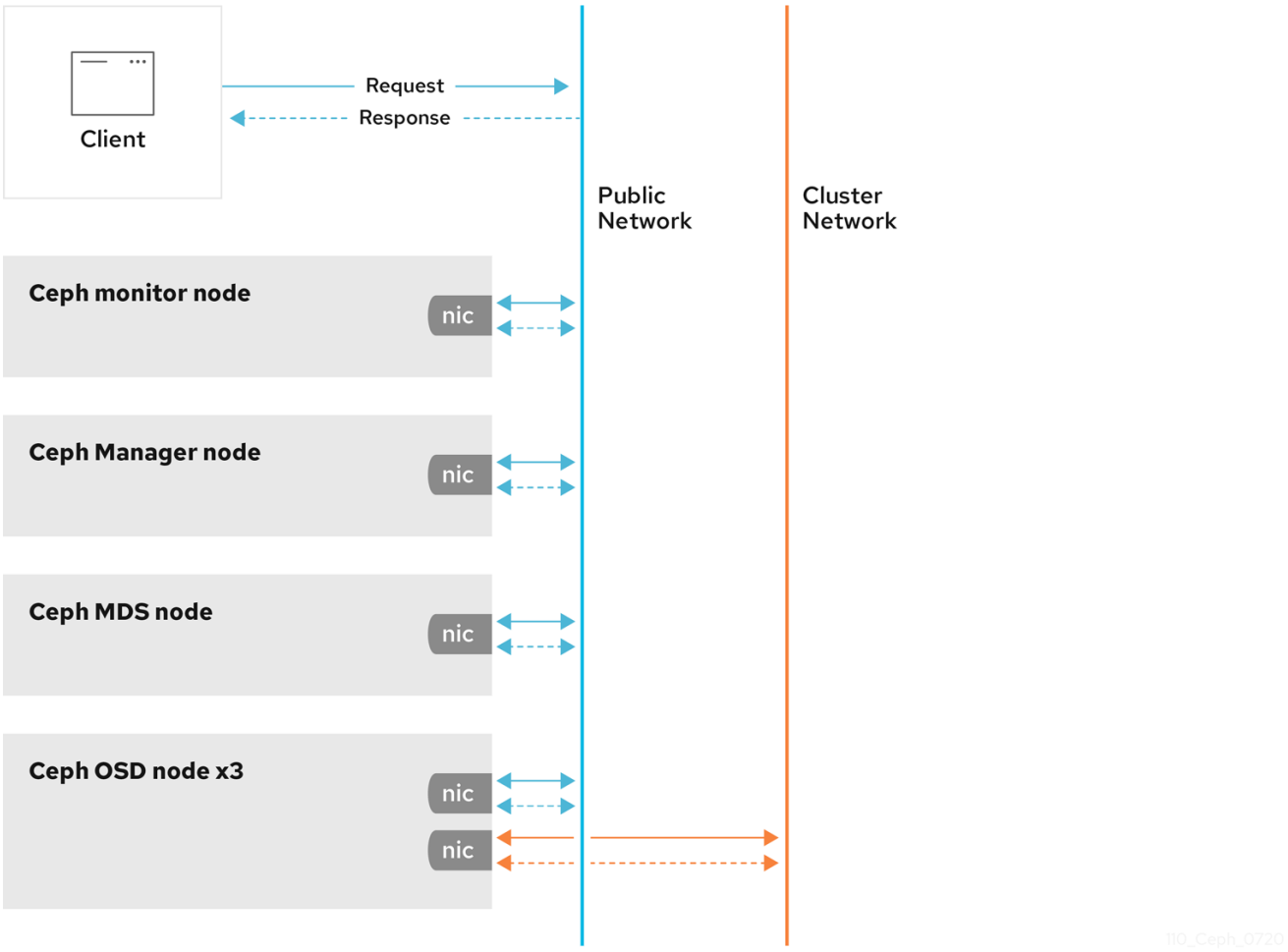
host 选项是节点的短名称，而不是 FQDN。它不是一个 IP 地址。

所有 Ceph 集群都必须使用公共网络。但是，除非指定了内部集群网络，Ceph 会假定有一个公共网络。Ceph 可以只使用公共网络运行，但对于大型存储集群，如果您有一个私有的专用网络用于处理与集群相关的网络流量，则性能会显著提升。

重要

红帽建议运行具有两个网络的 Ceph 存储集群。一个公共网络和一个专用网络。

要支持两个网络，每个 Ceph 节点都需要有一个以上的网络接口卡 (NIC) 。



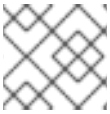
需要考虑操作两个独立网络的原因有很多：

- **性能：** Ceph OSD 处理 Ceph 客户端的数据复制。当 Ceph OSD 多次复制数据时，Ceph OSD 之间网络负载可轻松地在 Ceph 客户端和 Ceph 存储集群之间分配网络负载。这会引入延迟并创建性能问题。恢复和重新平衡还在公共网络上引入大量延迟。

- **安全** ：虽然大多数人用户都会正常使用资源，但有些人可能会参与所谓的拒绝服务 (DoS) 攻击。当 Ceph OSD 之间的流量中断时，peering 可能会失败，放置组可能无法反映 **active + clean** 状态，这可能会阻止用户读取和写入数据。缓解这类攻击的一个好方法是，维护一个完全独立的、不直接连接到互联网的集群网络。

不需要网络配置设置。Ceph 可只用于公共网络，假定在运行 Ceph 守护进程的所有主机上都配置了公共网络。但是，Ceph 允许您建立更加具体的标准，包括用于公共网络的多个 IP 网络和子网掩码。您还可以建立一个单独的集群网络来处理 OSD 心跳、对象复制和恢复流量。

不要将您在配置中设置的 IP 地址与面向公共的 IP 地址网络客户端混淆。典型的内部 IP 网络通常为 **192.168.0.0** 或 **10.0.0.0**。



注意

Ceph 使用 CIDR 表示法作为子网，如 **10.0.0.0/24**。

重要

如果您为公共或专用网络指定多个 IP 地址和子网掩码，则网络中的子网必须能够相互路由。另外，请确保在 IP 表中包括每个 IP 地址和子网，并根据需要打开端口。

当配置网络时，您可以重启集群或重启每个守护进程。Ceph 守护进程动态绑定，因此如果更改网络配置，不必一次重启整个集群。

其它资源

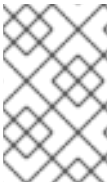
- 有关特定选项描述和使用，请参见 *Red Hat Ceph Storage 配置指南中的 附录 B*。

2.2. CEPH 网络 MESSENGER

messenger 是 Ceph 网络层实施。红帽支持两种 messenger 类型：

- **simple**
- **async**

在 Red Hat Ceph Storage 6 及更高版本中，**sync** 是默认的 mesenger 类型。要更改 messenger 类型，请在 Ceph 配置文件的 **[global]** 部分中指定 **ms_type** 配置设置。



注意

对于 **async** messenger，红帽支持 **posix** 传输类型，但目前不支持 **rdma** 或 **dpdk**。默认情况下，Red Hat Ceph Storage 中的 **ms_type** 设置反映 **async+posix**，其中 **async** 是 messenger 类型，**posix** 是传输类型。

SimpleMessenger

SimpleMessenger 实施使用每个套接字有两个线程的 TCP 套接字。Ceph 将每个逻辑会话与连接相关联。管道处理连接，包括每个消息的输入和输出。尽管 **SimpleMessenger** 对 **posix** 传输类型有效，但它不适用于 **rdma** 或 **dpdk** 等其他传输类型。

AsyncMessenger

因此，**AsyncMessenger** 是 Red Hat Ceph Storage 6 或更高版本的默认 mesenger 类型。对于 Red Hat Ceph Storage 6 或更高版本，**AsyncMessenger** 实现使用带有固定大小的线程池的 TCP 套接字进行连接，这应该等于最多副本数或纠删代码区块。如果性能因为 CPU 数量较低或每个服务器有大量 OSD，可以将线程数设置为较低值。

注意

红帽目前不支持其他传输类型，如 **rdma** 或 **dpdk**。

其它资源

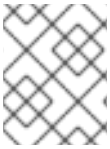
- 有关特定选项描述和使用，请参阅 *Red Hat Ceph Storage 配置指南 中的附录 B* AsyncMessenger 选项。
- 有关使用 Ceph messenger version 2 协议的 *on-wire encryption*，请参阅 *Red Hat Ceph Storage 架构指南*。

2.3. 配置公共网络

要配置 Ceph 网络，请在 **cephadm** shell 中使用 **config set** 命令。请注意，您在网络配置中设置的 IP 地址与网络客户端可能用来访问您的服务的面向公共 IP 地址不同。

Ceph 可只在一个公共网络中完美地正常工作。但是，Ceph 允许您建立更加具体的标准，包括用于公共网络的多个 IP 网络。

您还可以建立一个单独的私有集群网络来处理 OSD 心跳、对象复制和恢复流量。有关专用网络的更多信息，请参阅*配置专用网络*。



注意

Ceph 使用 CIDR 表示法作为子网，如 10.0.0.0/24。典型的内部 IP 网络通常是 192.168.0.0/24 或 10.0.0.0/24。

注意

如果您为公共或集群网络指定多个 IP 地址，则网络中的子网必须能够相互路由。另外，请确保在 IP 表中包括每个 IP 地址，并根据需要打开端口。

公共网络配置允许您为公共网络定义 IP 地址和子网。

先决条件

- 安装 Red Hat Ceph Storage 软件。

流程

1. 登录到 **cephadm** shell：

   示例

   ```
   [root@host01 ~]# cephadm shell
   ```

2. 使用子网配置公共网络：

   语法

   ```
   ceph config set mon public_network IP_ADDRESS_WITH_SUBNET
   ```

   示例

   ```
   [ceph: root@host01 /]# ceph config set mon public_network 192.168.0.0/24
   ```

3. 获取存储集群中的服务列表：

   示例

   ```
   [ceph: root@host01 /]# ceph orch ls
   ```

4. 重启守护进程。Ceph 守护进程动态绑定，因此如果更改特定守护进程的网络配置，不必一次重启整个集群。

   示例

   ```
   [ceph: root@host01 /]# ceph orch restart mon
   ```

5. 可选： 如果您要以 root 用户身份在 admin 节点上重启集群，请运行 **systemctl** 命令：

语法

```
systemctl restart ceph-FSID_OF_CLUSTER.target
```

示例

```
[root@host01 ~]# systemctl restart ceph-1ca9f6a8-d036-11ec-8263-fa163ee967ad.target
```

其它资源

- 了解具体选项说明和使用，请参阅 *Red Hat Ceph Storage 配置指南* 中的*附录 B*，。

2.4. 配置专用网络

不需要网络配置设置。Ceph 假设所有主机都在一个公共网络中，除非您专门配置集群网络，也称为*私有网络*。

如果您创建集群网络，OSD 会通过集群网络路由心跳、对象复制和恢复流量。与使用单个网络相比，这可以提高性能。

重要

为提高安全性，无法从公共网络或互联网访问集群网络。

要分配集群网络，请将 **--cluster-network** 选项与 **cephadm bootstrap** 命令一起使用。您指定的集群网络必须使用 CIDR 标记（如 10.90.90.0/24 or fe80::/64）定义子网。

您还可以在 boostrap 后配置 **cluster_network**。

先决条件

- 访问 Ceph 软件存储库。
- 对存储集群中所有节点的根级别访问权限。

流程

- 从您要用作存储集群中 monitor 节点的初始节点上运行 **cephadm bootstrap 命令**。在命令中包含 **--cluster-network** 选项。

语法

```
cephadm bootstrap --mon-ip IP-ADDRESS --registry-url registry.redhat.io --registry-username USER_NAME --registry-password PASSWORD --cluster-network NETWORK-IP-ADDRESS
```

示例

```
[root@host01 ~]# cephadm bootstrap --mon-ip 10.10.128.68 --registry-url registry.redhat.io --registry-username myuser1 --registry-password mypassword1 --cluster-network 10.10.0.0/24
```

- 要在 bootstrap 后配置 **cluster_network**，请运行 **config set** 命令并重新部署守护进程：

  1. 登录到 **cephadm** shell：

     示例

     ```
     [root@host01 ~]# cephadm shell
     ```

  2. 使用子网配置集群网络：

     语法

     ```
     ceph config set global cluster_network IP_ADDRESS_WITH_SUBNET
     ```

     示例

     ```
     [ceph: root@host01 /]# ceph config set global cluster_network 10.10.0.0/24
     ```

  3. 获取存储集群中的服务列表：

     示例

     ```
     [ceph: root@host01 /]# ceph orch ls
     ```

  4. 重启守护进程。Ceph 守护进程动态绑定，因此如果更改特定守护进程的网络配置，不必一次重启整个集群。

     示例

     ```
     [ceph: root@host01 /]# ceph orch restart mon
     ```

  5. 可选：如果您要以 root 用户身份在 admin 节点上重启集群，请运行 **systemctl** 命令：

     语法

     ```
     systemctl restart ceph-FSID_OF_CLUSTER.target
     ```

     示例

     ```
     [root@host01 ~]# systemctl restart ceph-1ca9f6a8-d036-11ec-8263-fa163ee967ad.target
     ```

其它资源

- 有关调用 **cephadm bootstrap** 的更多信息，请参阅 *Red Hat Ceph Storage 安装指南中的 Bootstrapping a new storage cluster* 部分。

2.5. 为集群配置多个公共网络

当用户想将 Ceph 监控守护进程放在属于多个网络子网的主机上时，需要配置多个公共网络到集群。用法示例是在 OpenShift Data Foundation 的 Metro DR 中用于 Advanced Cluster Management (ACM)的扩展集群模式。

您可以在 bootstrap 过程中将多个公共网络配置为集群，并在 bootstrap 完成后配置。

先决条件

- 在添加主机前，请确定您有一个正在运行的 Red Hat Ceph Storage 集群。

流程

1. 引导配置了多个公共网络的 Ceph 集群。
   a. 准备包含 **mon** 公共网络部分的 **ceph.conf** 文件：

重要

在当前用于 bootstrap 的主机上，必须至少配置一个提供的公共网络。

语法

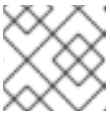
```
[mon]
public_network = PUBLIC_NETWORK1, PUBLIC_NETWORK2
```

示例

```
[mon]
public_network = 10.40.0.0/24, 10.41.0.0/24, 10.42.0.0/24
```

这是要为 bootstrap 提供的三个公共网络的示例。

   b. 通过提供 **ceph.conf** 文件作为输入来引导集群：

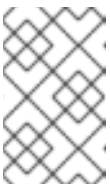


注意

在 bootstrap 中，您可以包含您要提供的任何其他参数。

语法

```
cephadm --image IMAGE_URL bootstrap --mon-ip MONITOR_IP -c
PATH_TO_CEPH_CONF
```



注意

或者，可以使用 ***IMAGE_ID***（如 **13ea90216d0be03003d12d7869f72ad9de5cec9e54a27fd308e01e467c0d4a0a**）替代 ***IMAGE_URL***。

示例

```
[root@host01 ~]# cephadm –image cp.icr.io/cp/ibm-ceph/ceph-5-rhel8:latest bootstrap –
mon-ip 10.40.0.0/24 -c /etc/ceph/ceph.conf
```

2. 在子网中添加新主机：

注意

正在添加的主机必须可从运行活跃管理器的主机访问。

a. 在新主机的 root 用户的 **authorized_keys** 文件中安装集群的公共 SSH 密钥：

语法

```
ssh-copy-id -f -i /etc/ceph/ceph.pub root@NEW_HOST
```

示例

```
[root@host01 ~]# ssh-copy-id -f -i /etc/ceph/ceph.pub root@host02
[root@host01 ~]# ssh-copy-id -f -i /etc/ceph/ceph.pub root@host03
```

b. 登录到 **cephadm** shell：

示例

```
[root@host01 ~]# cephadm shell
```

c. 将新主机添加到 Ceph 集群：

语法

```
ceph orch host add NEW_HOST IP [LABEL1 ...]
```

示例

```
[root@host01 ~]# ceph orch host add host02 10.10.0.102 label1
[root@host01 ~]# ceph orch host add host03 10.10.0.103 label2
```

注意

- 最好显式提供主机 IP 地址。如果没有提供 IP，则主机名会立即通过 DNS 解析，并使用该 IP。
- 也可以包含一个或多个标签来立即标记新主机。例如，默认情况下，**_admin** 标签使 cephadm 维护 **ceph.conf** 文件的副本，以及 **/etc/ceph** 目录中的 **client.admin** 密钥环文件。

3. 将公共网络参数的网络配置添加到正在运行的集群中。确保子网用逗号分开，且子网以 subnet/mask 格式列出。

语法

```
ceph config set mon public_network "SUBNET_1,SUBNET_2, ..."
```

示例

```
[root@host01 ~]# ceph config set mon public_network "192.168.0.0/24, 10.42.0.0/24, ..."
```

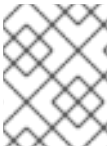
如有必要，更新 **mon** 规格，将 **mon** 守护进程放在指定子网中的主机上。

其它资源

- 如需有关在 *Red Hat Ceph Storage 安装指南中的* 添加主机的更多详细信息，请参阅添加主机。 *https://access.redhat.com/documentation/zh-cn/red_hat_ceph_storage/7/html-single/installation_guide/#adding-hosts_install*
- *如需了解扩展集群的更多信息，请参阅 Red Hat Ceph Storage Administration Guide 中的扩展集群部分。*

2.6. 验证为默认 CEPH 端口配置了防火墙规则

默认情况下，Red Hat Ceph Storage 守护进程使用 TCP 端口 6800-7100 与集群中的其他主机进行通信。您可以验证主机的防火墙是否允许连接这些端口。



注意

如果您的网络有一个专用防火墙，您可能需要在此流程之外验证其配置。如需更多信息，请参阅防火墙的文档。

如需更多信息，请参阅防火墙的文档。

先决条件

- *对主机的 root 级别访问。*

流程

1. *验证主机的 **iptables** 配置：*
   a. *列出活跃的规则：*

```
[root@host1 ~]# iptables -L
```

   b. *验证没有规则限制 TCP 端口 6800-7100 上的连接。*

   示例

```
REJECT all -- anywhere anywhere reject-with icmp-host-prohibited
```

2. *验证主机的 **firewalld** 配置：*
   a. *列出主机上打开的端口：*

   语法

```
firewall-cmd --zone ZONE --list-ports
```

   示例

```
[root@host1 ~]# firewall-cmd --zone default --list-ports
```

b. 验证范围是否为 TCP 端口 6800-7100。

2.7. CEPH 监控节点的防火墙设置

您可以通过引入 messenger 版本 2 协议，通过网络启用所有 Ceph 流量的加密。**me** senger v2 的安全模式设置加密 Ceph 守护进程和 Ceph 客户端之间的通信，从而为您提供端到端加密。

messenger v2 协议

Ceph on-wire 协议 **msgr2** 的第二个版本包括几个新功能：

- 安全模式可以加密通过网络移动的所有数据。
- 身份验证有效负载的封装改进。
- 功能公告和协商的改进。

Ceph 守护进程绑定到多个端口，允许旧旧 v1- 兼容新的 v2 兼容 Ceph 客户端，以连接同一存储集群。Ceph 客户端或其他 Ceph 守护进程连接到 Ceph Monitor 守护进程将先尝试使用 **v2** 协议（如果可能），但若不可能，则使用旧的 **v1** 协议。默认情况下，启用 messenger 协议 **v1** 和 **v2**。新的 v2 端口为 3300，旧的 v1 端口默认为 6789。

先决条件

- 一个正在运行的 Red Hat Ceph Storage 集群。
- 访问 Ceph 软件存储库。
- Ceph 监控节点的根级别访问权限.

流程

1. 使用以下示例添加规则：

```
[root@mon ~]# sudo iptables -A INPUT -i IFACE -p tcp -s IP-ADDRESS/NETMASK --dport
6789 -j ACCEPT
[root@mon ~]# sudo iptables -A INPUT -i IFACE -p tcp -s IP-ADDRESS/NETMASK --dport
3300 -j ACCEPT
```

   a. 将 **IFACE** 替换为公共网络接口（如 **eth0**、**eth1** 等等）。

   b. 将 **IP-ADDRESS** 替换为公共网络的 IP 地址，将 **NETMASK** 替换为公共网络的子网掩码。

2. 对于 **firewalld** 守护进程，执行以下命令：

```
[root@mon ~]# firewall-cmd --zone=public --add-port=6789/tcp
[root@mon ~]# firewall-cmd --zone=public --add-port=6789/tcp --permanent
[root@mon ~]# firewall-cmd --zone=public --add-port=3300/tcp
[root@mon ~]# firewall-cmd --zone=public --add-port=3300/tcp --permanent
```

其他资源

- 有关特定选项描述和使用，请参阅 Ceph 网络配置选项中的 Red Hat Ceph Storage 网络配置选项。

- *有关使用带有 Ceph messenger version 2 协议的 on-wire encryption，请参阅 Red Hat Ceph Storage 架构指南。*

第 3 章 CEPH 监控器配置

作为存储管理员，您可以使用 Ceph Monitor 的默认配置值，或根据预期的工作负载进行自定义。

先决条件

- 安装 Red Hat Ceph Storage 软件。

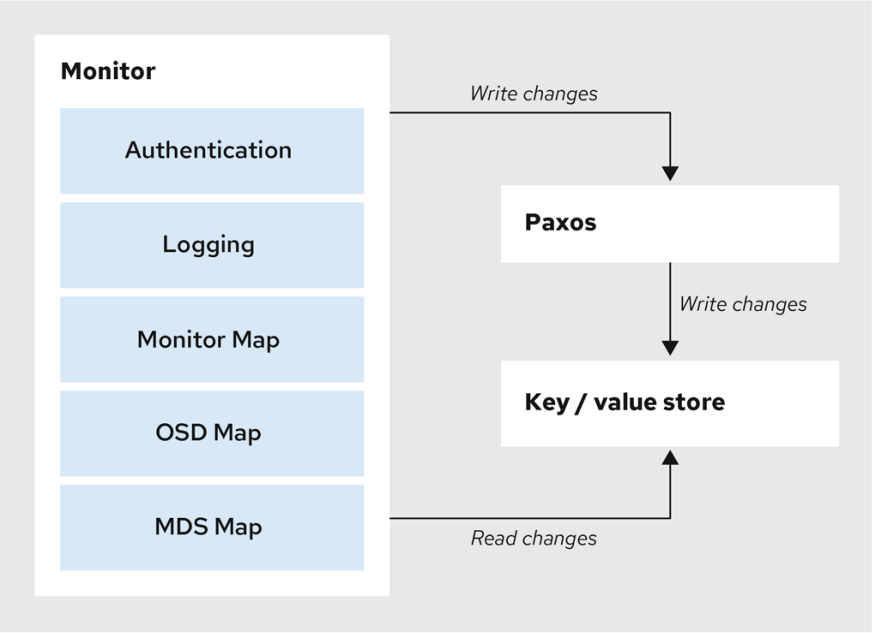
3.1. CEPH 监控器配置

了解如何配置 Ceph monitor 是构建可靠的 Red Hat Ceph Storage 集群的重要部分。所有存储集群都至少有一个监控器。Ceph 监控器配置通常保持一致性，但您可以在存储集群中添加、删除或替换 Ceph Monitor。

Ceph 监视器维护集群映射的"主副本"。这意味着，Ceph 客户端可以通过连接到一个 Ceph 监视器并检索当前 cluster map 来确定所有 Ceph 监视器和 Ceph OSD 的位置。

在 Ceph 客户端可以从 Ceph OSD 读取或写入到 Ceph OSD 之前，它们必须首先连接到 Ceph monitor。使用 cluster map 的当前副本和 CRUSH 算法时，Ceph 客户端可以计算任何对象的位置。计算对象位置的功能允许 Ceph 客户端直接与 Ceph OSD 通信，这是 Ceph 高可扩展性和性能的一个重要方面。

Ceph Monitor 的主要角色是维护集群映射的主副本。Ceph 监控程序也提供身份验证和日志记录服务。Ceph 监视器将监控服务中的所有更改写入单个 Paxos 实例，Paxos 会将更改写入到键值存储，以实现强一致性。Ceph monitor 可以在同步操作期间查询 cluster map 的最新版本。Ceph Monitor 利用键值存储的快照和迭代器（使用 **rocksdb** 数据库）来执行存储范围的同步。



3.2. 查看 CEPH 监控配置数据库

您可以在配置数据库中查看 Ceph Monitor 配置。

注意

以前的 Red Hat Ceph Storage 版本可把 Ceph Monitor 配置存储在 **/etc/ceph/ceph.conf** 中。从 Red Hat Ceph Storage 5 开始，这个配置文件已弃用。

先决条件

- 一个正在运行的 Red Hat Ceph Storage 集群。
- 对 Ceph 监控主机的 root 级别访问权限。

流程

1. 登录 **cephadm** shell。

```
[root@host01 ~]# cephadm shell
```

2. 使用 **ceph config** 命令查看配置数据库：

示例

```
[ceph: root@host01 /]# ceph config get mon
```

其它资源

- 有关 **ceph config** 命令可用的选项的更多信息，请使用 **ceph config -h**。

3.3. CEPH 集群映射

集群映射不同映射的一个组合，包括 monitor 映射、OSD 映射和放置组映射。集群映射跟踪多个重要事件：

- 哪些进程是 **in** Red Hat Ceph Storage 集群中。
- 哪些进程 **in** Red Hat Ceph Storage 集群处于 **up** 状态并运行或处于**down** 状态。
- 放置组是 **active** 或 **inactive**，以及 **clean** 或处于某些其他状态。
- 反映集群当前状态的其他详情，例如：
  - 存储空间总量或
  - 使用的存储量。

例如，当集群状态有重大改变时，Ceph OSD 会停机时，放置组将进入降级状态，以此类推。集群映射会更新，以反映集群的当前状态。此外，Ceph 监控器也维护了群集之前状态的历史记录。monitor 映射、OSD 映射和放置组映射各自维护其映射版本的历史记录。每个版本称为一个 **epoch**。

运行 Red Hat Ceph Storage 集群时，要跟踪这些状态是集群管理的一个重要部分。

3.4. CEPH MONITOR 仲裁

集群将通过单一监控器运行。但是，如果只有一个监控器，则代表有单一故障点。为确保生产 Ceph 存储群集中的高可用性，可运行具有多个监控器的 Ceph，因此当一个控制器出现故障时不会造成整个存储群集故障。

当 Ceph 存储集群运行多个 Ceph Monitor 以实现高可用性时，Ceph Monitor 使用 Paxos 算法来建立与主集群映射相关的共识。共识需要大多数运行的监控器建立一个仲裁（quorum）以达成对集群映射的共识。例如，1; 2 out of 3; 3 out of 5; 4 out of 6; 等

红帽建议运行至少有三个 Ceph Monitor 的生产环境 Red Hat Ceph Storage 集群，以确保高可用性。运行多个监视器时，您可以指定必须成为存储集群成员的初始监控器，才能建立仲裁。这可减少存储集群在线所需的时间。

```
[mon]
mon_initial_members = a,b,c
```

注意

存储集群中的大多数监控器都必须能够相互连接，以建立仲裁。您可以使用 **mon_initial_members** 选项减少初始 monitor 数量来建立仲裁。

3.5. CEPH MONITOR 一致性

将监控设置添加到 Ceph 配置文件时，您需要了解 Ceph 监控的一些架构方面。在发现集群中的另一个 Ceph monitor 时，Ceph 为 Ceph 监控器实施严格的一致性要求。Ceph 客户端和其他 Ceph 守护进程使用 Ceph 配置文件来发现 monitor，monitor 使用 monitor 映射(**monmap**)而不是 Ceph 配置文件互相发现。

在发现 Red Hat Ceph Storage 集群中的其他 Ceph 监控器时，Ceph Monitor 始终指 monitor map 的本地副本。使用 monitor 映射而不是 Ceph 配置文件可以避免会破坏集群的错误。例如，在指定监控地址或端口时 Ceph 配置文件中的拼写错误。由于 monitor 使用 monitor map 进行发现，并且它们与客户端和其他 Ceph 守护进程共享 monitor 映射，monitor 映射为 monitor 提供严格保证其共识有效。

将更新应用到 monitor 映射时会要求严格的一致性

与 Ceph monitor 上的任何其他更新一样，对 monitor 映射的更改始终通过名为 Paxos 的分布式共识算法运行。Ceph monitor 必须同意对 monitor 映射的每个更新，如添加或移除 Ceph monitor，以确保仲裁中的每个 monitor 都有相同的监控器映射版本。monitor 映射的更新是递增的，使 Ceph monitor 具有最新的商定版本和一组之前的版本。

维护历史记录

维护历史记录可让具有较老版本的 monitor 来获取 Red Hat Ceph Storage 集群的当前状态。

如果 Ceph 监控通过 Ceph 配置文件而不是监视器映射发现相互发现，它将带来额外的风险，因为 Ceph 配置文件没有被自动更新并分发。Ceph 监控可能会意外地使用旧的 Ceph 配置文件，无法识别 Ceph monitor，无法承担仲裁状态，或者开发 Paxos 无法准确确定系统当前状态的情况。

3.6. 引导 CEPH MONITOR

在大多数配置和部署情形中，部署 Ceph 的工具（如 **cephadm** ）可能会为您生成 monitor 来引导 Ceph 监视器。

Ceph 监控需要一些显式设置：

- **文件系统 ID ：fsid** 是对象存储的唯一标识符。由于您可以在同一硬件上运行多个存储集群，所以您必须在引导 monitor 时指定对象存储的唯一 ID。使用 **cephadm** 等部署工具会自动生成文件系统标识符，但您也可以手动指定 **fsid**。

- **Monitor ID** ： monitor ID 是分配给集群中各个 monitor 的唯一 ID。按照惯例，ID 设置为 monitor 的主机名。可以使用部署工具、**ceph** 命令或在 Ceph 配置文件中设置此选项。在 Ceph 配置文件中，按如下方式组成部分：

  示例

```
[mon.host1]

[mon.host2]
```

- **Key**: monitor 必须具有 secret 键。

其它资源

- 有关 **cephadm** 和 Ceph 编配器的更多信息，请参阅 Red Hat Ceph Storage Operations 指南 。

3.7. CEPH MONITOR 的最低配置

如果 Ceph 配置文件中尚未配置 DNS 和 monitor 地址，则 Ceph Monitor 的裸机监控设置包括每个 monitor 的主机名。默认情况下，Ceph 监控在端口 **6789** 和 **3300** 上运行。

重要

不要编辑 Ceph 配置文件。

注意

此 monitor 的最小配置假定部署工具为您生成 **fsid** 和 **mon.** 键。

您可以使用以下命令设置或读取存储集群配置选项。

- **ceph config dump** 整个存储集群配置数据库。
- **Ceph config generate-minimal-conf** - 生成最小 **ceph.conf** 文件。
- **ceph config get WHO** - 转储特定守护进程或客户端的配置，如 Ceph 监控器配置数据库中存储。
- **ceph config set WHO OPTION VALUE** - 在 Ceph 监控配置数据库中设置配置选项。
- **ceph config show WHO** - 显示所报告的运行中守护进程配置。
- **ceph config assimilate-conf -i INPUT_FILE -o OUTPUT_FILE** - 用于从输入文件中获取配置文件，并将任何有效选项移到 Ceph Monitor 的配置数据库中。

这里，WHO 参数可以是项的名称或一个 Ceph 守护进程， OPTION 是一个配置文件，VALUE 可以是 **true** 或 **false**。

重要

当 Ceph 守护进程在从 config 存储获取选项前需要配置选项时，您可以通过运行以下命令来设置配置：

```
ceph cephadm set-extra-ceph-conf
```

此命令将文本添加到所有守护进程的 **ceph.conf** 文件中。它是一个临时解决方案，不是推荐的操作。

3.8. CEPH 的唯一标识符

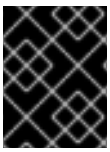
每个 Red Hat Ceph Storage 集群都有一个唯一标识符（**fsid**）。如果指定，它通常出现在配置文件的 **[global]** 部分中。部署工具通常会生成 **fsid** 并将其存储在 monitor 映射中，因此该值可能不会出现在配置文件中。通过 **fsid**，可以在同一硬件上为多个集群运行守护进程。

注意

如果使用部署工具，则不要设置这个值。

3.9. CEPH MONITOR 数据存储

Ceph 提供了 Ceph 监视器存储数据的默认路径。



重要

红帽建议在独立于 Ceph OSD 的驱动器中运行 Ceph 监视器，以便在生产环境 Red Hat Ceph Storage 集群中获得最佳性能。

注意

专用 **/var/lib/ceph** 分区应该用于 MON 数据库，大小介于 50 到 100 GB 之间。

Ceph 监控器经常调用 **fsync()** 函数，这可能会影响 Ceph OSD 工作负载。

Ceph 监视器将其数据存储为键值对。使用数据存储可防止恢复 Ceph 监视器通过 Paxos 运行损坏版本，而且它可在一个原子批处理中实现多次修改操作，以及其他优势。

重要

红帽不推荐修改默认数据位置。如果您修改默认位置，请通过在配置文件的 **[mon]** 部分中设置它，使它在 Ceph 监视器间统一。

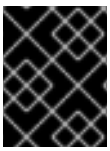
3.10. CEPH 存储容量

当 Red Hat Ceph Storage 集群接近其最大容量时（通过 **mon_osd_full_ratio** 参数显示），Ceph 会阻止您写入或读取 Ceph OSD 的安全措施，以防止数据丢失。因此，使一个生产环境的 Red Hat Ceph Storage 集群接近其全满比率是一个不好的做法，因为它降低了高可用性。默认全满比率为 **.95** 或 95% 的容量。对于具有多个 OSD 的测试集群来说，这是一个非常积极的设置。

提示

监控集群时，请注意与 **nearfull** 比率相关的警告。这意味着，一些 OSD 故障可能会导致临时服务中断，如果一个或多个 OSD 出现故障。考虑添加更多 OSD 以增加存储容量。

测试集群的常见场景涉及系统管理员从 Red Hat Ceph Storage 集群中删除 Ceph OSD，以观察集群重新平衡。然后，删除另一个 Ceph OSD，直到 Red Hat Ceph Storage 集群最终达到完全的比例并锁定。



重要

红帽建议在一个测试集群中仍然有点的容量规划。通过规划，您可以量化您需要的备用容量来保持高可用性。

理想情况下，您要规划一系列 Ceph OSD 失败的情况，集群可以在不立即替换这些 Ceph OSD 的情况下恢复到 **active + clean** 状态。您可以运行状态为 **active + degraded** 的集群，但这对正常操作并不是一个理想的状态。

下图显示了一个简化的 Red Hat Ceph Storage 集群，其中包含每个主机有一个 Ceph OSD 的 33 Ceph 节点，每个 Ceph OSD 守护进程从中读取并写入 3TB 驱动器。因此，这一 exemplary Red Hat Ceph Storage 集群具有最大 99TB 的实际容量。当 **mon osd full ratio** 为 **0.95**，如果 Red Hat Ceph Storage 集群达到 5TB 的容量，集群不允许 Ceph 客户端读取和写入数据。因此，Red Hat Ceph Storage 集群的操作容量为 95 TB，而不是 99 TB。

110_Ceph_0720

在这样的集群中，一个或多个 OSD 无法正常使用。较为频繁但合理的方案涉及机架的路由器或电源故障，例如同时导致多个 OSD 下线，例如 OSDs 7-12。在这种情况下，保持集群正常运行并处于 **active + clean** 状态仍会为您带来更大益处，即使这需要在短时间内添加具有额外 OSD 的主机。如果您的容量利用率太高，可能不会丢失数据，但您仍然可能会牺牲数据可用性，同时在故障域内解决集群的容量利用率超过完整的比例。因此，红帽建议至少使用一些最小容量规划。

识别集群的两个值：

- OSD 数量
- 集群的总容量

要确定集群中的 OSD 的平均容量，请将集群的总容量除以集群中的 OSD 数量。考虑将这个数量乘以您希望在正常操作期间同时出现故障的 OSD 数量（相对较小的数）。最后，通过满比例将集群的容量乘以达到最大操作容量。然后，从 OSD 中减去您希望无法达到合理的全满比率的 OSD 的数据量。重复处理数量较高的 OSD 故障（例如，一个 OSD 机架），以达到接近的全满比率的合理数量。

3.11. CEPH 心跳

Ceph 监视器通过要求来自每个 OSD 的报告以及从 OSD 接收关于其邻居 OSD 状态的报告来了解集群。Ceph 为 monitor 和 OSD 之间的交互提供了合理的默认设置，但您可以根据需要修改它们。

3.12. CEPH MONITOR 同步角色

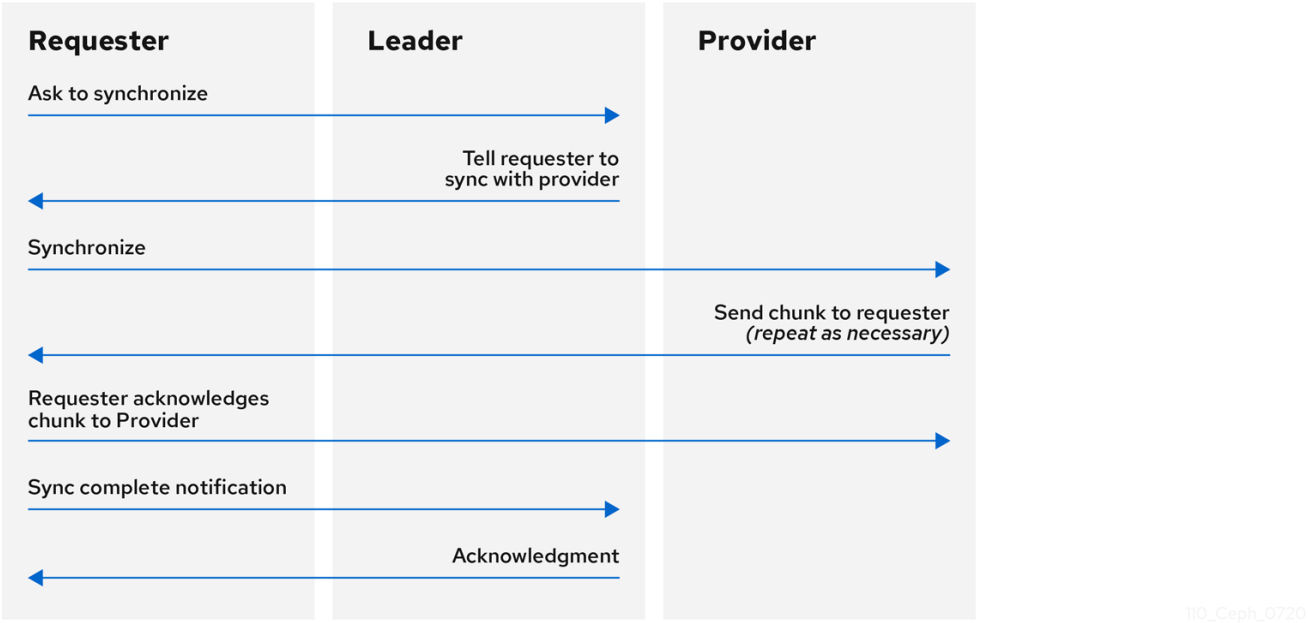
当您使用多个监控器运行生产集群时，每个 monitor 会检查邻居监视器是否有最新版本的 cluster map。例如，一个邻居监视器中的映射，其一个或多个 epoch 号高于即时监视器映射中当前 epoch 的数值。定期，集群中的一个监视器可能位于其他 monitor 后，它必须离开仲裁，同步来检索有关集群的最新信息，然后重新加入仲裁。

同步角色

出于同步目的，监视器可以假定以下三个角色之一：

- **Leader** ：Leader 是达到集群映射的最新 Paxos 版本的第一个 moitor。
- **Provider** ：Provider 是一个具有集群映射的最新版本的 monitor，但不是第一个。
- **Requester:** 请求者是一个监控器，它已落后于领导，必须同步来检索集群的最新信息，然后才能重新加入仲裁。

这些角色使领导能够将同步任务委派给提供程序，从而防止同步请求过载，并提高性能。在下图中，请求者已了解到它已位于其他 monitor 后。请求者要求领导要同步，并且领导者告诉请求者与提供程序同步。



监控同步

当新 monitor 加入集群时，才会发生同步。在运行时操作期间，监控器可以在不同时间接收集群映射的的更新。这意味着领导和提供商角色可以从一个监控器迁移到另一个监视器。例如在同步时发生这种情况，例如，提供商落于领导，提供商可以与请求者终止同步。

同步完成后，Ceph 需要在集群中修剪。修剪要求放置组处于 **active + clean** 状态。

3.13. CEPH 时间同步

Ceph 守护进程将关键消息传递到彼此，这必须在守护进程到达超时阈值前进行处理。如果 Ceph 监视器中的时钟没有同步，则可以导致一些异常。

例如：

- 忽略过期时间戳等消息的守护进程。
- 当没有收到信息时，超时会马上或晚末触发。

提示

在 Ceph 监控主机上安装 NTP，以确保监控集群与时钟同步运行。

时钟偏移可能仍然可以通过 NTP 发现，即使差异尚未有害。Ceph 时钟偏移和时钟偏移警告可能会触发，即使 NTP 维护合理的同步级别。在这种情况下，可以容忍时钟偏移。但是，很多因素，如工作负载、网络延迟、配置为默认超时，其他同步选项会影响可接受的时钟偏移级别，而不影响 Paxos 保证。

其它资源

- 如需了解更多详细信息，请参阅 Ceph 时间同步 部分。
- 有关特定选项描述和使用，请参阅 Ceph Monitor 配置选项中的所有 Red Hat Ceph Storage Monitor 配置选项。

第 4 章 CEPH 身份验证配置

作为存储管理员，对用户和服务进行身份验证对于 Red Hat Ceph Storage 集群的安全性至关重要。Red Hat Ceph Storage 包含 Cephx 协议，作为加密身份验证的默认协议，以及管理存储集群中的身份验证的工具。

Red Hat Ceph Storage 包含 Cephx 协议，作为加密身份验证的默认协议，以及管理存储集群中的身份验证的工具。

作为 Ceph 身份验证配置的一部分，请考虑 Ceph 和网关守护进程的密钥轮转，以提高安全性。密钥轮转是通过命令行使用 **cephadm** 进行的。如需了解更多详细信息，请参阅启用密钥轮转。

先决条件

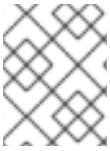
- 安装 Red Hat Ceph Storage 软件。

4.1. CEPHX 身份验证

cephx 协议默认启用。加密身份验证具有一些计算成本，尽管它们通常很低。如果连接了客户端和服务器的网络环境被视为安全性，并且您无法负担身份验证计算成本，您可以禁用它。在部署 Ceph 存储集群时，部署工具将创建 **client.admin** 用户和密钥环。

重要

红帽建议使用身份验证。



注意

如果您禁用身份验证，您将面临中间攻击的风险，改变客户端和服务器信息，这可能会导致严重的安全问题。

启用和禁用 Cephx

启用 Cephx 要求您已部署 Ceph 监控器和 OSD 的密钥。在切换 Cephx 身份验证的打开和关闭时，不必重复部署步骤。

4.2. 启用 CEPHX

启用 **cephx** 后，Ceph 将在默认搜索路径中查找密钥环，其中包括 **/etc/ceph/$cluster.$name.keyring**。您可以通过在 Ceph 配置文件的 **[global]** 部分添加 **keyring** 选项来覆盖该位置，但不建议这样做。

执行以下步骤，在禁用身份验证的集群中启用 **cephx**。如果您或部署实用程序生成了密钥，您可以跳过与生成密钥相关的步骤。

先决条件

- 一个正在运行的 Red Hat Ceph Storage 集群。
- Ceph 监控节点的根级别访问权限.

流程

1. 创建 **client.admin** 密钥，并为您的客户端主机保存密钥副本：
   -

```
[root@mon ~]# ceph auth get-or-create client.admin mon 'allow *' osd 'allow *' -o
/etc/ceph/ceph.client.admin.keyring
```

警告

这将擦除任何现有 **/etc/ceph/client.admin.keyring** 文件的内容。如果部署工具已为您完成，则不要执行此步骤。

2. 为 monitor 集群创建密钥环，并生成监控器 secret 密钥：

```
[root@mon ~]# ceph-authtool --create-keyring /tmp/ceph.mon.keyring --gen-key -n mon. --
cap mon 'allow *'
```

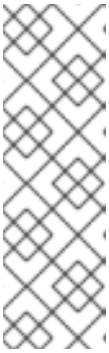
3. 将 monitor keyring 复制到每个 monitor **mon data** 目录中的 **ceph.mon.keyring** 文件。例如，要将其复制到集群 **ceph** 中的 **mon.a** 中，请使用：

```
[root@mon ~]# cp /tmp/ceph.mon.keyring /var/lib/ceph/mon/ceph-a/keyring
```

4. 为每个 OSD 生成 secret 密钥，其中 **ID** 是 OSD 号：

```
ceph auth get-or-create osd.ID mon 'allow rwx' osd 'allow *' -o
/var/lib/ceph/osd/ceph-ID/keyring
```

5. 默认情况下启用 **cephx** 身份验证协议。



注意

如果在以前通过将身份验证选项设置为 **none** 禁用了 **cephx** 身份验证协议，那么删除 Ceph 配置文件 (**/etc/ceph/ceph.conf**) 中的 **[global]** 部分的以下行来重新启用 **cephx** 身份验证协议：

```
auth_cluster_required = none
auth_service_required = none
auth_client_required = none
```

6. 启动或重启 Ceph 存储集群。

重要

启用 **cephx** 需要停机，因为集群需要完全重启，或者在禁用客户端 I/O 时将其关闭并启动。

这些标记需要在重启或关闭存储集群前设置：

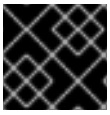
```
[root@mon ~]# ceph osd set noout
[root@mon ~]# ceph osd set norecover
[root@mon ~]# ceph osd set norebalance
[root@mon ~]# ceph osd set nobackfill
[root@mon ~]# ceph osd set nodown
[root@mon ~]# ceph osd set pause
```

启用 **cephx** 后，所有 PG 都活跃且干净，取消设置标记：

```
[root@mon ~]# ceph osd unset noout
[root@mon ~]# ceph osd unset norecover
[root@mon ~]# ceph osd unset norebalance
[root@mon ~]# ceph osd unset nobackfill
[root@mon ~]# ceph osd unset nodown
[root@mon ~]# ceph osd unset pause
```

4.3. 禁用 CEPHX

以下流程描述了如何禁用 Cephx。如果您的集群环境相对安全，您可以降低运行身份验证的计算费用。



重要

红帽建议启用身份验证。

但是，在设置或故障排除过程中可能会更容易地禁用身份验证。

先决条件

- 一个正在运行的 Red Hat Ceph Storage 集群。
- Ceph 监控节点的根级别访问权限.

流程

1. 通过在 Ceph 配置文件的 **[global]** 部分设置以下选项来禁用 **cephx** 身份验证：

   示例

   ```
   auth_cluster_required = none
   auth_service_required = none
   auth_client_required = none
   ```

2. 启动或重启 Ceph 存储集群。

4.4. CEPHX 用户密钥环

当您运行 Ceph 并启用了身份验证时，**ceph** 管理命令和 Ceph 客户端需要身份验证密钥来访问 Ceph 存储集群。

向 **ceph** 管理命令和客户端提供这些密钥的最常见方式是在 **/etc/ceph/** 目录下包含 Ceph 密钥环。文件名通常是 **ceph.client.admin.keyring** 或 **$cluster.client.admin.keyring**。如果您在 **/etc/ceph/** 目录下包含密钥环，则不需要在 Ceph 配置文件中指定 **keyring** 条目。

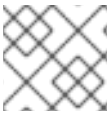
重要

红帽建议将 Red Hat Ceph Storage 集群 keyring 文件复制到您要运行管理命令的节点，因为它包含 **client.admin** 密钥。

要做到这一点，请执行以下命令：

```
# scp USER@HOSTNAME:/etc/ceph/ceph.client.admin.keyring /etc/ceph/ceph.client.admin.keyring
```

将 **USER** 替换为主机上的用户名，使用 **client.admin** 键，将 **HOSTNAME** 替换为该主机的主机名。



注意

确保 **ceph.keyring** 文件已在客户端计算机上设置适当的权限。

您可以使用 **key** 设置在 Ceph 配置文件中（不推荐）指定密钥本身，或使用 **keyfile** 设置来指定到密钥文件的路径。

4.5. CEPHX 守护进程密钥环

管理用户或部署工具可能会像生成用户密钥环一样生成守护进程密钥环。默认情况下，Ceph 将守护进程密钥环存储在其数据目录中。默认密钥环位置，以及守护进程正常工作所需的功能。

注意

monitor 密钥环包含密钥，但没有功能，不是 Ceph 存储集群 **auth** 数据库的一部分。

守护进程数据目录位置默认为表单的目录：

```
/var/lib/ceph/$type/CLUSTER-ID
```

示例

```
/var/lib/ceph/osd/ceph-12
```

您可以覆盖这些位置，但不推荐进行。

4.6. CEPHX 消息签名

Ceph 提供精细的控制，您可以为客户端和 Ceph 之间的服务消息启用或禁用签名。您可以为 Ceph 守护进程之间的消息启用或禁用签名。

重要

红帽建议 Ceph 使用为该初始身份验证设置的会话密钥验证实体之间的所有持续消息。

注意

Ceph 内核模块尚不支持签名。

第 5 章 池、放置组和 CRUSH 配置

作为存储管理员，您可以选择将 Red Hat Ceph Storage 默认选项用于池、放置组和 CRUSH 算法，或者为预期的工作负载自定义它们。

先决条件

- 安装 Red Hat Ceph Storage 软件。

5.1. 池放置组和 CRUSH

当您创建池并为池设置放置组数量时，Ceph 在不特别覆盖默认值时会使用默认值。

重要

红帽建议覆盖一些默认值。特别是，设置池的副本大小并覆盖默认放置组数量。

您可以在运行池命令时设置这些值。

默认情况下，Ceph 生成 3 个对象副本。如果要将对象的 4 个副本设置为默认值，一个主和三个副本，重新设置默认值，如 **osd_pool_default_size** 所示。如果要允许 Ceph 在降级状态写入副本数，将 **osd_pool_default_min_size** 设置为比 **osd_pool_default_size** 值小的数字。

示例

```
[ceph: root@host01 /]# ceph config set global osd_pool_default_size 4  # Write an object 4 times.
[ceph: root@host01 /]# ceph config set global osd_pool_default_min_size 1  # Allow writing one copy in a degraded state.
```

确定您有实际的放置组数量。红帽建议每个 OSD 大约 100 个。例如，OSD 的总数乘以 100 的倍数并除以副本数（**osd_pool_default_size**）。对于 10 个 OSD 和 **osd_pool_default_size** = 4，我们建议的值为 (100 * 10)/ 4 = 250。

示例

```
[ceph: root@host01 /]# ceph config set global osd_pool_default_pg_num 250
[ceph: root@host01 /]# ceph config set global osd_pool_default_pgp_num 250
```

其他资源

- 有关特定选项描述和使用，请参阅 IAppendix E 中的所有 Red Hat Ceph Storage 池、放置组和 CRUSH 配置选项。

第 6 章 CEPH OBJECT STORAGE DAEMON (OSD) 配置

作为存储管理员，您可以将 Ceph Object Storage Daemon (OSD) 配置为基于预期的工作负载冗余和优化。

先决条件

- 安装 Red Hat Ceph Storage 软件。

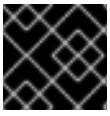
6.1. CEPH OSD 配置

所有 Ceph 集群都有一个配置，它定义：

- 集群身份
- 身份验证设置
- 集群中的 Ceph 守护进程成员资格
- 网络配置
- 主机名和地址
- keyring 的路径
- 到 OSD 日志文件的路径
- 其他运行时选项

部署工具（如 **cephadm** ）通常会为您创建初始 Ceph 配置文件。但是，如果您想在不使用部署工具的情况下引导集群，您可以自己创建一个。

为方便起见，每个守护进程都有一系列默认值。许多由 **ceph/src/common/config_opts.h** 脚本设置。您可以通过 Ceph 配置文件或运行时覆盖这些设置，方法是使用 monitor **tell** 命令，或直接连接到 Ceph 节点上的守护进程套接字。



重要

红帽不推荐更改默认路径，因为以后对 Ceph 进行故障排除更困难。

其它资源

- 有关 **cephadm** 和 Ceph 编配器的更多信息，请参阅 Red Hat Ceph Storage Operations 指南 。

6.2. 刮除 OSD

除了生成多个对象副本外，Ceph 还能通过清理放置组来确保数据完整性。Ceph 清理与对象存储层上的 **fsck** 命令类似。

对于每个放置组，Ceph 都会生成所有对象的目录，并比较每个主对象及其副本，以确保缺少对象或不匹配。

轻度清理（每日）会检查对象大小和属性。深度刮除（每周）读取数据并使用 checksum 来确保数据完整性。

清理对于保持数据完整性非常重要，但可能会降低性能。调整以下设置以增加或减少清理操作。

其他资源

- 如需了解更多详细信息，请参阅 Red Hat Ceph Storage 配置指南 中的 Ceph 清理选项。

6.3. 回填 OSD

将 Ceph OSD 添加到集群或从集群中删除时，CRUSH 算法会通过将放置组移到 Ceph OSD 或从中移出来重新平衡集群。迁移放置组和包含的对象可以大大降低集群操作性能。为保持操作性能，Ceph 会使用"回填"进程执行此迁移，这使 Ceph 将回填操作设置为比读取或写入数据的请求较低优先级。

6.4. OSD 恢复

当集群启动或 Ceph OSD 意外终止并重启时，OSD 在出现写入操作前开始与其他 Ceph OSD 的对等。

如果 Ceph OSD 崩溃然后又恢复在线，通常它将与其他 Ceph OSD 同步，包含 PG 中最新版本的对象。发生这种情况时，Ceph OSD 进入恢复模式并寻求数据的最新副本，并使其映射重新变为最新。根据 Ceph OSD 停机的时长，OSD 对象和放置组可能会显著不同步。另外，如果故障域停止（例如，一个机架出现问题），则在恢复过程中可能出现多个 Ceph OSD 同时上线的问题。这样可使恢复过程消耗和大量资源。

为保持可操作的性能，Ceph 对恢复请求、线程数和对象块大小（允许 Ceph 处于降级状态）执行恢复。

其他资源

- 有关特定选项描述和使用，请参阅 OSD 对象守护进程存储配置选项 中的所有 Red Hat Ceph Storage Ceph OSD 配置选项。

第 7 章 CEPH 监控和 OSD 交互配置

作为存储管理员，您必须正确配置 Ceph 监控器和 OSD 之间的交互，以确保稳定的工作环境。

先决条件

- 安装 Red Hat Ceph Storage 软件。

7.1. CEPH 监控和 OSD 交互

完成初始 Ceph 配置后，您可以部署并运行 Ceph。当您执行 **ceph health** 或 **ceph -s** 等命令时，Ceph Monitor 会报告 Ceph Storage 集群的当前状态。Ceph Monitor 通过从每个 Ceph OSD 守护进程要求报告来了解 Ceph 存储集群，并通过从 Ceph OSD 守护进程接收报告来了解其邻居 Ceph OSD 守护进程的状态。如果 Ceph Monitor 没有接收报告，或者它收到 Ceph 存储集群中更改的报告，Ceph 监控器会更新 Ceph 集群映射的状态。

Ceph 为 Ceph Monitor 和 OSD 交互提供合理的默认设置。但是，您可以覆盖默认值。以下小节论述了 Ceph 监控器和 Ceph OSD 守护进程如何进行交互，以满足监控 Ceph 存储集群的目的。

7.2. OSD 心跳

每个 Ceph OSD 守护进程会每 6 秒检查其他 Ceph OSD 守护进程的心跳。要更改心跳间隔，请在运行时更改值：

语法

```
ceph config set osd osd_heartbeat_interval TIME_IN_SECONDS
```

示例

```
[ceph: root@host01 /]# ceph config set osd osd_heartbeat_interval 60
```

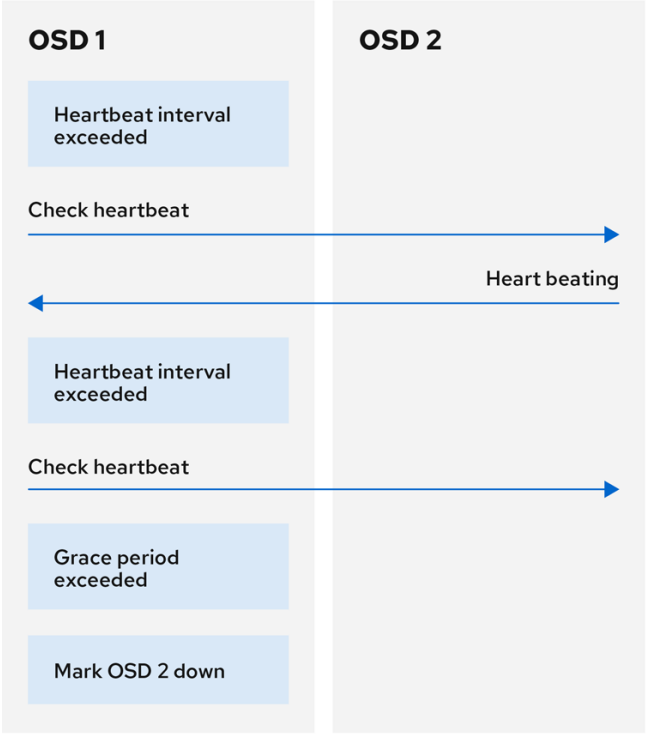
如果邻居 Ceph OSD 守护进程没有在 20 秒宽限期内发送心跳数据包，Ceph OSD 守护进程可能会认为邻居 Ceph OSD 守护进程 **down**。它可以报告回 Ceph monitor，它会更新 Ceph cluster map。要更改宽限期，请在运行时设置值：

语法

```
ceph config set osd osd_heartbeat_grace TIME_IN_SECONDS
```

示例

```
[ceph: root@host01 /]# ceph config set osd osd_heartbeat_grace 30
```



7.3. 将 OSD 报告为 DOWN

默认情况下，**来自不同主机**的两个 Ceph OSD 守护进程必须报告给另一个 Ceph OSD 守护进程处于 **down** 状态的 Ceph 监控器，然后确认报告的 Ceph OSD 守护进程为 **down**。

但是，所有 OSD 报告失败的可能性都位于具有错误交换机的机架中，导致 OSD 之间的连接问题。

为避免"错误警报"，Ceph 会将故障报告为类似 lagy 的"subcluster"的代理。虽然情况并非总是如此，但可能帮助管理员对性能不良的系统子集进行本地化处理。

Ceph 使用 **mon_osd_reporter_subtree_level** 设置，将 peer 分到"cluster"的常用级别类型。

默认情况下，仅需要两个**来自不同子树**的报告，才能报告另一个 Ceph OSD 守护进程为 **down**。管理员可以通过在运行时设置 **mon_osd_min_down_reporters** 和 **mon_osd_reporter_subtree_level** 的值，修改报告者的数量，将用于报告一个 Ceph OSD Daemon **down** 所需的唯一的子树和祖先类型改为 Ceph Monitor：

语法

```
ceph config set mon mon_osd_min_down_reporters NUMBER
```

示例

```
[ceph: root@host01 /]# ceph config set mon mon_osd_min_down_reporters 4
```

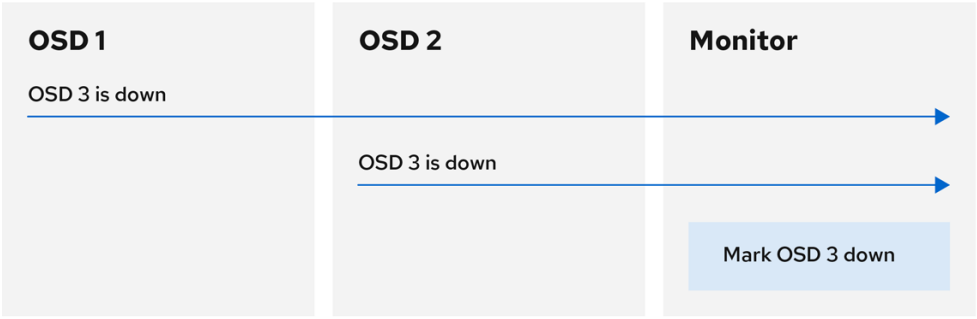
语法

```
ceph config set mon mon_osd_reporter_subtree_level CRUSH_ITEM
```

示例

-

```
[ceph: root@host01 /]# ceph config set mon mon_osd_reporter_subtree_level host
[ceph: root@host01 /]# ceph config set mon mon_osd_reporter_subtree_level rack
[ceph: root@host01 /]# ceph config set mon mon_osd_reporter_subtree_level osd
```



7.4. 报告对等故障

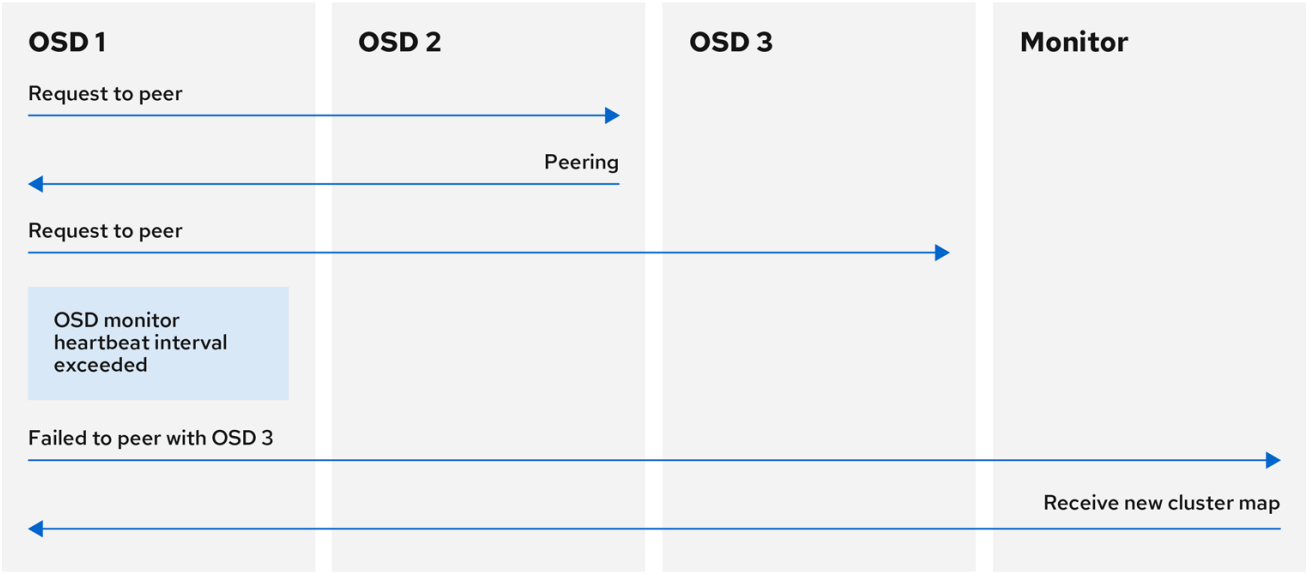
如果 Ceph OSD 守护进程无法与其 Ceph 配置文件或 cluster map 中定义的任何 Ceph OSD 守护进程的对等点，它会每 30 秒对集群 map 的最新副本发出 Ceph Monitor 命令。您可以通过在运行时设置值来更改 Ceph 监控心跳间隔：

语法

```
ceph config set osd osd_mon_heartbeat_interval TIME_IN_SECONDS
```

示例

```
[ceph: root@host01 /]# ceph config set osd osd_mon_heartbeat_interval 60
```



7.5. OSD 报告状态

如果 Ceph OSD 守护进程没有报告到 Ceph 监控器，Ceph Monitor **会在 mon_osd_report_timeout** 之后标记 Ceph OSD 守护进程，即 900 秒。当可报告事件（如故障）时，Ceph OSD 守护进程会向 Ceph 监控器发送报告，这是放置组统计的变化、**up_thru** 或在 5 秒内引导时发生的变化。

您可以通过在运行时设置 **osd_mon_report_interval** 值来更改 Ceph OSD 守护进程最小报告间隔：

语法

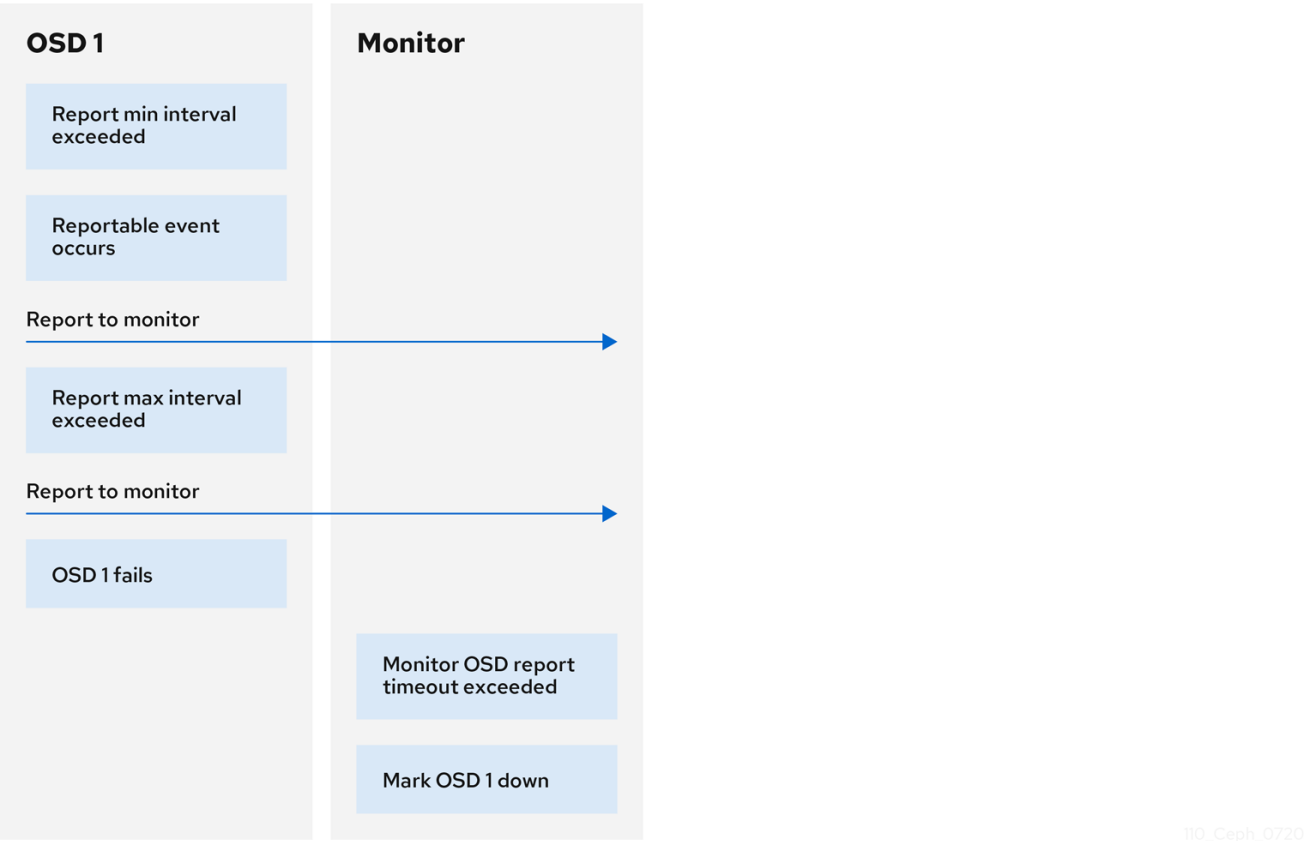
```
ceph config set osd osd_mon_report_interval TIME_IN_SECONDS
```

要获取、设置和验证配置，您可以使用以下示例：

示例

```
[ceph: root@host01 /]# ceph config get osd osd_mon_report_interval
5
[ceph: root@host01 /]# ceph config set osd osd_mon_report_interval 20
[ceph: root@host01 /]# ceph config dump | grep osd

global          advanced  osd_pool_default_crush_rule          -1
  osd           basic     osd_memory_target                    4294967296
  osd           advanced  osd_mon_report_interval              20
```



其他资源

- 有关特定选项描述和使用，请参阅 Ceph Monitor 和 OSD 配置选项中的所有 Red Hat Ceph Storage Ceph Monitor 和 OSD 配置选项。

第 8 章 CEPH 调试和日志记录配置

作为存储管理员，您可以增加 **cephadm** 中的调试和日志记录信息，以帮助诊断 Red Hat Ceph Storage 的问题。

先决条件

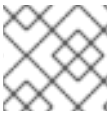
- 安装了 Red Hat Ceph Storage 软件。

其他资源

- 有关特定选项描述和使用，请参阅 Ceph 调试和日志配置选项中的所有 Red Hat Ceph Storage Ceph 调试和日志记录配置选项。
- 有关 **cephadm** 故障排除的更多信息，请参阅 Red Hat Ceph Storage Administration Guide 中的 Cephadm 故障排除。
- 有关 **cephadm** 日志记录的更多信息，请参阅 Red Hat Ceph Storage 管理指南中的 Cephadm 操作。

附录 A. 常规配置选项

这些是 Ceph 的一般配置选项。



注意

通常，它们将通过部署工具（如 **cephadm** ）自动设置。

fsid

描述

文件系统 ID。每个集群一个。

类型

UUID

必填

No.

默认

N/A.通常由部署工具生成。

admin_socket

描述

在守护进程上执行管理命令的套接字，无论 Ceph 监视器是否建立了仲裁。

类型

字符串

必填

否

默认

/var/run/ceph/$cluster-$name.asok

pid_file

描述

监控或 OSD 将在其中写入其 PID 的文件。例如，**/var/run/$cluster/$type.$id.pid** 将为在 **ceph** 集群中运行的 id 为 **a** 的 **mon** 创建 /var/run/ceph/mon.a.pid。当守护进程安全停止时，将删除 **pid 文件**。如果进程不是守护进程化（使用 **-f** 或 **-d** 选项运行），则不会创建 **pid 文件**。

类型

字符串

必填

否

默认

否

chdir

描述

目录 Ceph 守护进程在启动并运行后会变为。建议使用默认 / 目录。

类型

字符串

必填

否

默认

/

max_open_files

描述

如果设置，当 Red Hat Ceph Storage 集群启动时，Ceph 会在 OS 级别设置 ***max_open_fds*** *（即，最大文件描述符的数量）。它有助于防止 Ceph OSD 耗尽文件描述符。*

类型

64 位整数

必填

否

默认

0

fatal_signal_handlers

描述

如果设置，我们将为 SEGV、ABRT、BUS、ILL、FPE、XCPU、XCPU、XZ、SYS 信号安装信号处理器，以生成有用的日志消息。

类型

布尔值

默认

true

附录 B. CEPH 网络配置选项

这些是 Ceph 的常见网络配置选项。

public_network

描述

公共（前端）网络的 IP 地址和子网掩码（例如，**192.168.0.0/24**）。在 **[global]** 中设置。您可以指定以逗号分隔的子网。

类型

<ip-address>/<netmask> [, <ip-address>/<netmask>]

必需

否

默认

N/A

public_addr

描述

公共（前端）网络的 IP 地址。为每个守护进程设置。

类型

IP 地址

必填

否

默认

N/A

cluster_network

描述

集群网络的 IP 地址和子网掩码（例如 **10.0.0.0/24**）。在 **[global]** 中设置。您可以指定以逗号分隔的子网。

类型

<ip-address>/<netmask> [, <ip-address>/<netmask>]

必需

否

默认

N/A

cluster_addr

描述

集群网络的 IP 地址。为每个守护进程设置。

类型

地址

必需

否

默认

N/A

ms_type

描述

*网络传输层的 messenger 类型。红帽支持使用 **posix** 语义的 **simple** 和 **async** messenger 类型。*

类型

字符串.

必填

No.

默认

async+posix

ms_public_type

描述

*公共网络的网络传输层的 messenger 类型。它的工作方式与 **ms_type** 相同，但仅适用于公共网络或前端网络。此设置可让 Ceph 为公共或前端和集群或后端网络使用不同的 mesenger 类型。*

类型

字符串.

必填

No.

默认

无。

ms_cluster_type

描述

*集群网络的网络传输层的 messenger 类型。它的工作方式与 **ms_type** 相同，但仅适用于集群或后端网络。此设置可让 Ceph 为公共或前端和集群或后端网络使用不同的 mesenger 类型。*

类型

字符串.

必填

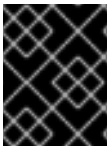
No.

默认

无。

主机选项

*您必须在 Ceph 配置文件中至少声明一个 Ceph Monitor，每个声明的 monitor 下都有一个 **mon addr** 设置。Ceph 预期 Ceph 在 Ceph 配置文件中每个声明的 monitor、元数据服务器和 OSD 下的 **host** 设置。*



重要

*不要使用 **localhost**。使用节点的短名称，而不是完全限定域名 (FQDN) 。在使用检索节点名称的第三方部署系统时，不要为 **host** 指定任何值。*

mon_addr

描述

客户端可用于连接到 Ceph 监视器的 **<hostname>:<port>** 条目列表。如果没有设置，Ceph 会搜索 **[mon.*]** 部分。

类型

字符串

必填

否

默认

N/A

主机

描述

主机名。对特定的守护进程实例使用此设置（例如，**[osd.0]**）。

类型

字符串

必需

是，对于守护进程实例。

默认

localhost

TCP 选项

Ceph 默认禁用 TCP 缓冲。

ms_tcp_nodelay

描述

Ceph 启用 **ms_tcp_nodelay**，使每个请求立即发送（无缓冲区）。禁用 Nagle 的算法会增加网络流量，它可以造成拥塞问题。如果您遇到大量小数据包，您可以尝试禁用 **ms_tcp_nodelay**，但请注意，禁用它通常会增加延迟。

类型

布尔值

必需

否

默认

true

ms_tcp_rcvbuf

描述

网络连接结尾的套接字缓冲区的大小。默认禁用此选项。

类型

32 位整数

必填

否

默认

0

ms_tcp_read_timeout

描述

如果客户端或守护进程向另一个 Ceph 守护进程发出请求且不丢弃未使用的连接，则 **tcp read timeout** 会在指定秒数后将连接定义为 idle。

类型

unsigned 64 位整数

必填

否

默认

900 15 分钟。

绑定选项

bind 选项配置 Ceph OSD 守护进程的默认端口范围。默认范围为 **6800:7100**。您还可以启用 Ceph 守护进程来绑定到 IPv6 地址。

重要

验证防火墙配置是否允许您使用配置的端口范围。

ms_bind_port_min

描述

OSD 守护进程要绑定到的最低端口号。

类型

32 位整数

默认

6800

必填

否

ms_bind_port_max

描述

OSD 守护进程要绑定到的最大端口号。

类型

32 位整数

默认

7300

必填

No.

ms_bind_ipv6

描述

启用 Ceph 守护进程来绑定到 IPv6 地址。

类型

布尔值

默认

false

必需

否

异步 messenger 选项

这些 Ceph mesenger 选项配置 **AsyncMessenger** 的行为。

ms_async_transport_type

描述

AsyncMessenger 使用的传输类型。红帽支持 **posix** 设置，但目前不支持 **dpdk** 或 **rdma** 设置。POSIX 使用标准 TCP/IP 网络，它是默认值。其他传输类型是实验性的，因此**不被**支持。

类型

字符串

必填

否

默认

posix

ms_async_op_threads

描述

每个 **AsyncMessenger** 实例使用的初始 worker 线程数。此配置设置 **SHOULD** 等于副本或纠删代码区块的数量，但如果 CPU 内核数量较低或单一服务器上的 OSD 数量很高，则可设置它。

类型

64-bit Unsigned 整数

必需

否

默认

3

ms_async_max_op_threads

描述

每个 **AsyncMessenger** 实例使用的最大 worker 线程数量。如果 OSD 主机具有有限 CPU 数量，并且如果 Ceph 利用率不足，则设置为较低值。

类型

64-bit Unsigned 整数

必需

否

默认

5

ms_async_set_affinity

描述

设置为 ***true***，*将* ***AsyncMessenger*** *worker 绑定到特定的 CPU 内核。*

类型

布尔值

必需

否

默认

true

ms_async_affinity_cores

描述

当 ***ms_async_set_affinity*** *为* ***true*** *时，该字符串指定了将* ***AsyncMessenger*** *worker 绑定到 CPU 内核的方式。例如：* ***0,2*** *会将 worker #1 和 #2 分别绑定到 CPU 内核 #0 和 #2。* ***注意：*** *在手动设置关联性时，确保不将 worker 分配给创建的虚拟 CPU，作为超线程或类似技术的影响，因为它们比物理 CPU 内核慢。*

类型

字符串

必填

否

默认

(empty)

ms_async_send_inline

描述

直接从生成它们的线程中直接发送消息，而不是排队并从 ***AsyncMessenger*** *线程发送发送。这个选项已知可以降低具有大量 CPU 内核的系统的性能，因此默认禁用它。*

类型

布尔值

必需

否

默认

false

连接模式配置选项

对于大多数连接，有控制用于加密和解密的模式的选项。

ms_cluster_mode

描述

用于 Ceph 守护进程之间的集群内通信的连接模式。如果列出了多个模式，则首选列出模式。

类型

字符串

默认

CRC secure

ms_service_mode

描述

连接到存储集群时要使用的允许模式列表。

类型

字符串

默认

CRC secure

ms_client_mode

描述

连接模式列表（按首选顺序），供客户端在与 Ceph 集群交互时使用。

类型

字符串

默认

CRC secure

ms_mon_cluster_mode

描述

在 Ceph 监视器之间使用的连接模式。

类型

字符串

默认

Secure crc

ms_mon_service_mode

描述

客户端或其他 Ceph 守护进程在连接 monitor 时要使用的允许模式列表。

类型

字符串

默认

Secure crc

ms_mon_client_mode

描述

连接模式列表（按首选顺序），用于连接到 Ceph 监视器时使用的客户端或非监视器守护进程。

类型

字符串

默认

Secure crc

压缩模式配置选项

使用 messenger v2 协议，您可以使用压缩模式的配置选项。

ms_compress_secure

描述

将加密与压缩相结合可降低对等点之间的消息安全程度。如果同时启用了加密和压缩，则忽略压缩设置，且不会压缩消息。使用 选项覆盖此设置。直接从生成它们的线程中直接发送消息，而不是排队并从 **AsyncMessenger** 线程发送发送。这个选项已知可以降低具有大量 CPU 内核的系统的性能，因此默认禁用它。

类型

布尔值

默认

false

ms_osd_compress_mode

描述

消息传递中用于与 Ceph OSD 通信的压缩策略。

类型

字符串

默认

none

有效选择

none 或 **force**

ms_osd_compress_min_size

描述

最少的消息大小适用于在线压缩。

类型

整数

默认

1 Ki

ms_osd_compression_algorithm

描述

与 OSD 连接的压缩算法按首选顺序

类型

字符串

默认

snappy

有效选择

snappy,**zstd**,**zlib**, 或 **lz4**

附录 C. CEPH 监控配置选项

以下是部署期间可以设置的 Ceph 监控配置选项。

您可以使用 **ceph config set mon CONFIGURATION_OPTION VALUE** 命令设置这些配置选项。

mon_initial_members

描述

启动期间集群中初始 monitor 的 ID。如果指定，Ceph 需要奇数个 monitor 来形成初始仲裁（例如 3）。

类型

字符串

默认

无

mon_force_quorum_join

描述

强制 monitor 加入仲裁，即使之前已从映射中删除

类型

布尔值

默认

False

mon_dns_srv_name

描述

用于查询监控主机/地址的 DNS 的服务名称。

类型

字符串

默认

ceph-mon

fsid

描述

集群 ID。每个集群一个。

类型

UUID

必填

是。

默认

N/A.如果未指定，则由部署工具生成。

mon_data

描述

monitor 的数据位置。

类型

字符串

默认

/var/lib/ceph/mon/$cluster-$id

mon_data_size_warn

描述

当监控数据存储达到这个阈值时，Ceph 在集群日志记录中发出 ***HEALTH_WARN*** *状态。默认值为 15GB。*

类型

整数

默认

15*1024*1024*1024

mon_data_avail_warn

描述

当 monitor 数据存储的可用磁盘空间低于此百分比时，Ceph 会在集群日志记录中发出 ***HEALTH_WARN*** *状态。*

类型

整数

默认

30

mon_data_avail_crit

描述

当 monitor 数据存储的可用磁盘空间较低或等于这个百分比时，Ceph 会在集群日志记录中发出 ***HEALTH_ERR*** *状态。*

类型

整数

默认

5

mon_warn_on_cache_pools_without_hit_sets

描述

如果缓存池没有设置 ***hit_set_type*** *参数，Ceph 会在集群日志记录中发出* ***HEALTH_WARN*** *状态。*

类型

布尔值

默认

true

mon_warn_on_crush_straw_calc_version_zero

描述

如果 CRUSH 的 ***straw_calc_version*** *为零，Ceph 会在集群日志记录中发出* ***HEALTH_WARN*** *状态。详情请参阅 CRUSH 可调项。*

类型

布尔值

默认

true

mon_warn_on_legacy_crush_tunables

描述

如果 CRUSH 可调项太旧（相对于 **mon_min_crush_required_version**而言，Ceph 会在集群日志记录中发出 **HEALTH_WARN** 状态）。

类型

布尔值

默认

true

mon_crush_min_required_version

描述

此设置定义集群所需的最小可调配置集版本。

类型

字符串

默认

hammer

mon_warn_on_osd_down_out_interval_zero

描述

如果 **mon_osd_down_out_interval** 设置为零，Ceph 在集群日志中会发出 **HEALTH_WARN** 状态，因为设置了 **noout** 标记时 Leader 的行为也类似。管理员通过设置 **noout** 标志来更轻松地对集群进行故障排除。Ceph 发出警告，以确保管理员知道该设置为零。

类型

布尔值

默认

true

mon_cache_target_full_warn_ratio

描述

当处于 **cache_target_full** 和 **target_max_object** 比率之间时，Ceph 会发出警告。

类型

浮点值

默认

0.66

mon_health_data_update_interval

描述

仲裁（以秒为单位）监控器与其对等状态共享其健康状况。负数会禁用运行状况更新。

类型

浮点值

默认

60

mon_health_to_clog

描述

此设置可让 Ceph 定期向集群日志发送健康摘要。

类型

布尔值

默认

true

mon_health_detail_to_clog

描述

此设置可让 Ceph 定期向集群日志记录发送健康详情。

类型

布尔值

默认

true

mon_op_complaint_time

描述

在没有更新后，Ceph Monitor 操作被视为被阻断的秒数。

类型

整数

默认

30

mon_health_to_clog_tick_interval

描述

monitor 将健康摘要发送到集群日志记录的频率（以秒为单位）。一个非正数代表禁用。如果当前的健康摘要为空，或者与上一次相同，则 monitor 不会将状态发送到集群日志。

类型

整数

默认

60.000000

mon_health_to_clog_interval

描述

monitor 将健康摘要发送到集群日志记录的频率（以秒为单位）。一个非正数代表禁用。该监控将始终向群集日志发送摘要。

类型

整数

默认

600

mon_osd_full_ratio

描述

在 OSD 被视为 **full** 之前使用的磁盘空间百分比。

类型

浮点值：

默认

.95

mon_osd_nearfull_ratio

描述

在 OSD 视为 **nearfull** 之前使用的磁盘空间百分比。

类型

浮点值

默认

.85

mon_sync_trim_timeout

描述, 类型

双

默认

30.0

mon_sync_heartbeat_timeout

描述, 类型

双

默认

30.0

mon_sync_heartbeat_interval

描述, 类型

双

默认

5.0

mon_sync_backoff_timeout

描述, 类型

双

默认

30.0

mon_sync_timeout

描述

监控器从同步提供程序等待下一次更新消息的秒数，然后再再次提供引导。

类型

双

默认

60.000000

mon_sync_max_retries

描述, 类型

整数

默认

5

mon_sync_max_payload_size

描述

同步有效负载（以字节为单位）的最大大小。

类型

32 位整数

默认

1045676

paxos_max_join_drift

描述

必须首先同步监控数据存储前，最大的 Paxos 迭代。当 monitor 发现其对等点比其太超前时，它将首先与数据存储同步，然后再继续。

类型

整数

默认

10

paxos_stash_full_interval

描述

（在 commits 时）stash PaxosService 状态的完整副本的频率。目前，此设置仅影响 **mds**、**mon**、**auth** 和 **mgr** PaxosServices。

类型

整数

默认

25

paxos_propose_interval

描述

收集这个时间更新，然后再执行映射更新。

类型

双

默认

1.0

paxos_min

描述

要保留的最小 paxos 状态数量

类型

整数

默认

500

paxos_min_wait

描述

在不活跃的一段时间后，收集更新的最小时间。

类型

双

默认

0.05

paxos_trim_min

描述

在修剪前可以容忍的额外提议数

类型

整数

默认

250

paxos_trim_max

描述

一次要修剪的最大额外提议数

类型

整数

默认

500

paxos_service_trim_min

描述

触发修剪的最小版本数量（0 禁用它）

类型

整数

默认

250

paxos_service_trim_max

描述

单一提案期间要修剪的最大版本量（0 代表禁用它）

类型

整数

默认

500

mon_max_log_epochs

描述

单个提议期间要修剪的最大日志时期量

类型

整数

默认

500

mon_max_pgmap_epochs

描述

单个建议期间要修剪的最大 pgmap epoch 数量

类型

整数

默认

500

mon_mds_force_trim_to

描述

强制 monitor 在这点上修剪 mdsmaps（0 代表禁用。这个设置比较危险，请谨慎使用）

类型

整数

默认

0

mon_osd_force_trim_to

描述

强制 monitor 在这点上修剪 osdmaps，即使指定 epoch 中没有清理 PG（0 则禁用它。dangerous 则谨慎使用）

类型

整数

默认

0

mon_osd_cache_size

描述

osdmaps 缓存的大小，不依赖于底层存储的缓存

类型

整数

默认

500

mon_election_timeout

描述

在选举代理上，让所有 ACK 的最大等待时间（以秒为单位）。

类型

浮点值

默认

5

mon_lease

描述

监控版本中租期的长度（以秒为单位）。

类型

浮点值

默认

5

mon_lease_renew_interval_factor

描述

***mon lease* * *mon lease renew interval factor* 将是领导机更新其他 monitor 的租期的时间间隔。因素应小于 *1.0*。**

类型

浮点值

默认

0.6

mon_lease_ack_timeout_factor

描述

*领导机将会等待 **mon lease** * **mon lease ack timeout factor** 的时间来等待供应商确认租期的扩展。*

类型

浮点值

默认

2.0

mon_accept_timeout_factor

描述

*领导机将等待 **mon lease** * **mon accept timeout** 来等待 Requester 接受 Paxos 更新。它还在 Paxos 恢复阶段用于类似目的。*

类型

浮点值

默认

2.0

mon_min_osdmap_epochs

描述

始终保留的最小 OSD map epoch 数。

类型

32 位整数

默认

500

mon_max_pgmap_epochs

描述

监视器应保留的最大 PG 映射 epoch 数。

类型

32 位整数

默认

500

mon_max_log_epochs

描述

监视器应保留的最大日志 epoch 数。

类型

32 位整数

默认

500

clock_offset

描述

系统时钟偏移量。详情请查看 ***Clock.cc****。*

类型

双

默认

0

mon_tick_interval

描述

监视器的空循环间隔（以秒为单位）。

类型

32 位整数

默认

5

mon_clock_drift_allowed

描述

监视器之间允许的时钟偏移（以秒为单位）。

类型

浮点值

默认

.050

mon_clock_drift_warn_backoff

描述

用于时钟偏移警告的指数 backoff.

类型

浮点值

默认

5

mon_timecheck_interval

描述

检查领导的时间间隔（时钟偏移检查）。

类型

浮点值

默认

300.0

mon_timecheck_skew_interval

描述

在领导存在偏差时（以秒为单位）的时间检查间隔（时钟偏移检查）。

类型

浮点值

默认

30.0

mon_max_osd

描述

集群中允许的最大 OSD 数量。

类型

32 位整数

默认

10000

mon_globalid_prealloc

描述

为集群中的客户端和守护进程预先分配的全局 ID 数量。

类型

32 位整数

默认

10000

mon_sync_fs_threshold

描述

在编写指定对象数量时与文件系统同步。将它设置为 ***0*** *以禁用它。*

类型

32 位整数

默认

5

mon_subscribe_interval

描述

订阅的刷新间隔（以秒为单位）。订阅机制允许获取集群映射和日志信息。

类型

双

默认

86400.000000

mon_stat_smooth_intervals

描述

Ceph 将在最后的 ***N*** *PG 映射中平稳统计信息。*

类型

整数

默认

6

mon_probe_timeout

描述

监视器在 bootstrap 前等待的对等点的秒数。

类型

双

默认

2.0

mon_daemon_bytes

描述

存储服务器和 OSD 消息的消息内存大写（以字节为单位）。

类型

64 位 Unsigned 整数

默认

400ul << 20

mon_max_log_entries_per_event

描述

每个事件的最大日志条目数。

类型

整数

默认

4096

mon_osd_prime_pg_temp

描述

当 OSD 返回到集群中时，启用或禁用带有之前 OSD 的 PGMap。使用 **true** 设置时，客户端将继续使用前面的 OSD，直到新的 OSD 已作为 PG 的对等。

类型

布尔值

默认

true

mon_osd_prime_pg_temp_max_time

描述

当 OSD 返回到集群时，monitor 应该花费多少时间（以秒为单位）尝试推断 PGMap。

类型

浮点值

默认

0.5

mon_osd_prime_pg_temp_max_time_estimate

描述

在我们并行控制所有 PG 前，每个 PG 花费的最大估算时间。

类型

浮点值

默认

0.25

mon_osd_allow_primary_affinity

描述

允许在 osdmap 中设置 **primary_affinity**。

类型

布尔值

默认

False

mon_osd_pool_ec_fast_read

描述

是否启用对池的快速读取。如果在创建时没有指定 ***fast_read****，它将用作新创建的池的默认设置。*

类型

布尔值

默认

False

mon_mds_skip_sanity

描述

如果出现错误，则跳过 FSMap 上的安全断言。如果 FSMap sanity 检查失败，则 monitor 会终止，但您可以通过启用此选项来禁用它。

类型

布尔值

默认

False

mon_max_mdsmap_epochs

描述

单一建议期间要修剪的最大 mdsmap epoch 数。

类型

整数

默认

500

mon_config_key_max_entry_size

描述

config-key 条目的最大数量（以字节为单位）。

类型

整数

默认

65536

mon_warn_pg_not_scrubbed_ratio

描述

scrub max interval 的百分比超过 scrub max interval to warn。

类型

浮点值

默认

0.5

mon_warn_pg_not_deep_scrubbed_ratio

描述

深度清理间隔的百分比超过深度清理间隔以警告

类型

浮点值

默认

0.75

mon_scrub_interval

描述

monitor 将存储清理其存储的频率（以秒为单位）与所有存储的密钥的计算方式进行比较。

类型

整数

默认

*3600*24*

mon_scrub_timeout

描述

重启 mon 仲裁参与者清理的超时时间不会响应最新的块。

类型

整数

默认

5 分钟

mon_scrub_max_keys

描述

每次清理的最大键数。

类型

整数

默认

100

mon_scrub_inject_crc_mismatch

描述

注入 CRC 不匹配的可能性到 Ceph Monitor scrub 中。

类型

整数

默认

*3600*24*

mon_scrub_inject_missing_keys

描述

将缺少的密钥注入到 mon scrub 中的概率。

类型

浮点值

默认

0

mon_compact_on_start

描述

*在 **ceph-mon** 启动时，紧凑用作 Ceph Monitor 存储的数据库。手动压缩有助于缩小 monitor 数据库，并在常规压缩失败时提高其性能。*

类型

布尔值

默认

False

mon_compact_on_bootstrap

描述

在 bootstrap 时紧凑用作 Ceph Monitor 存储的数据库。monitor 会相互探测到 bootstrap 后创建仲裁。如果在加入仲裁前超时，它将再次启动并引导自身。

类型

布尔值

默认

False

mon_compact_on_trim

描述

在我们修剪旧状态时，压缩特定的前缀（包括 paxos）。

类型

布尔值

默认

true

mon_cpu_threads

描述

在监控中执行 CPU 密集型工作的线程数量。

类型

布尔值

默认

true

mon_osd_mapping_pgs_per_chunk

描述

我们以块的形式计算从放置组到 OSD 的映射。这个选项指定每个块的放置组数量。

类型

整数

默认

4096

mon_osd_max_split_count

描述

每个 "involved" OSD 的最大 PG 数量，以便进行拆分创建。当增加池的 **pg_num** *时，放置组将划分到为这个池的所有 OSD 上。我们希望避免在 PG 分上的极倍。*

类型

整数

默认

300

rados_mon_op_timeout

描述

从监控器返回错误前等待监视器的响应秒数。0 表示无限，或者没有等待时间。

类型

双

默认

0

其它资源

- *池值*
- *CRUSH 可调项*

附录 D. CEPHX 配置选项

以下是部署期间可以设置的 Cephx 配置选项。

auth_cluster_required

描述

如果启用，Red Hat Ceph Storage 集群守护进程 **ceph-mon** 和 **ceph-osd** 必须相互进行身份验证。有效的设置是 **cephx** 或 **none**。

类型

字符串

必填

否

默认

cephx。

auth_service_required

描述

如果启用，Red Hat Ceph Storage 集群守护进程需要 Ceph 客户端与 Red Hat Ceph Storage 集群进行身份验证，以便能访问 Ceph 服务。有效的设置是 **cephx** 或 **none**。

类型

字符串

必填

否

默认

cephx。

auth_client_required

描述

如果启用，Ceph 客户端需要 Red Hat Ceph Storage 集群与 Ceph 客户端进行身份验证。有效的设置是 **cephx** 或 **none**。

类型

字符串

必填

否

默认

cephx。

keyring

描述

keyring 文件的路径。

类型

字符串

必填

否

默认

/etc/ceph/$cluster.$name.keyring,/etc/ceph/$cluster.keyring,/etc/ceph/keyring,/etc/ceph/keyring.bin

keyfile

描述

密钥文件的路径（即仅包含密钥的文件）。

类型

字符串

必填

否

默认

无

key

描述

密钥（即密钥本身，它是一个文本字符串）。不建议。

类型

字符串

必填

否

默认

无

ceph-mon

位置

$mon_data/keyring

功能

mon 'allow *'

ceph-osd

位置

$osd_data/keyring

功能

mon 'allow profile osd' osd 'allow *'

radosgw

位置

$rgw_data/keyring

功能

mon 'allow rwx' osd 'allow rwx'

cephx_require_signatures

描述

如果设置为 **true**，Ceph 需要在 Ceph 客户端和 Red Hat Ceph Storage 集群间的所有消息流量上签名，并且守护进程之间由 Red Hat Ceph Storage 集群组成。

类型

布尔值

必需

否

默认

false

cephx_cluster_require_signatures

描述

如果设置为 **true**，Ceph 需要 Ceph 守护进程间的所有消息流量签名，由 Red Hat Ceph Storage 集群组成。

类型

布尔值

必需

否

默认

false

cephx_service_require_signatures

描述

如果设置为 **true**，Ceph 需要 Ceph 客户端和 Red Hat Ceph Storage 集群间的所有消息流量签名。

类型

布尔值

必需

否

默认

false

cephx_sign_messages

描述

如果 Ceph 版本支持消息签名，Ceph 会对所有消息进行签名，使其无法欺骗。

类型

布尔值

默认

true

auth_service_ticket_ttl

描述

当 Red Hat Ceph Storage 集群向 Ceph 客户端发送一个 ticket 进行身份验证时，集群会为 ticket 分配一个生存时间。

类型

双

默认

60*60

附录 E. 池、放置组和 CRUSH 配置选项

管理池、放置组和 CRUSH 算法的 Ceph 选项。

mon_allow_pool_delete

描述

允许 monitor 删除池。在 RHCS 3 及更新的版本中，监控器无法默认删除池，以保护数据。

类型

布尔值

默认

false

mon_max_pool_pg_num

描述

每个池的最大放置组数量。

类型

整数

默认

65536

mon_pg_create_interval

描述

在同一 Ceph OSD 守护进程中创建 PG 间隔的秒数。

类型

浮点值

默认

30.0

mon_pg_stuck_threshold

描述

PG 被视为卡住的秒数。

类型

32 位整数

默认

300

mon_pg_min_inactive

描述

如果处于非活跃状态的时间超过 **mon_pg_stuck_threshold** 的 PG 数量超过这个设置，Ceph 会在集群日志记录中记录一个 **HEALTH_ERR** 状态。默认设置为一个 PG。非正数代表禁用此设置。

类型

整数

默认

1

mon_pg_warn_min_per_osd

描述

如果集群中每个 OSD 的平均 PG 数量小于此设置，Ceph 会在集群日志记录中发出 ***HEALTH_WARN*** *状态。非正数代表禁用此设置。*

类型

整数

默认

30

mon_pg_warn_max_per_osd

描述

如果集群中每个 OSD 的平均 PG 数量大于此设置，Ceph 会在集群日志记录中发出 ***HEALTH_WARN*** *状态。非正数代表禁用此设置。*

类型

整数

默认

300

mon_pg_warn_min_objects

描述

如果集群中的对象总数低于这个数字，则不发出警告。

类型

整数

默认

1000

mon_pg_warn_min_pool_objects

描述

不要在对象数低于这个值的池中警告。

类型

整数

默认

1000

mon_pg_check_down_all_threshold

描述

down *OSD 的阈值（百分比），在超过这个值时 Ceph 检查所有 PG 以确保它们没有处于 stuck 或 stale 状态。*

类型

浮点值

默认

0.5

mon_pg_warn_max_object_skew

描述

如果池中对象平均数量大于 **mon pg warn max object skew** 乘以所有池的平均数量，则 Ceph 在集群日志中发出 **HEALTH_WARN** 状态。非正数代表禁用此设置。

类型

浮点值

默认

10

mon_delta_reset_interval

描述

在 Ceph 将 PG 增量重置为零之前需要经过的不活跃的秒数。Ceph 追踪各个池的已用空间增量，以帮助管理员评估恢复和性能的进度。

类型

整数

默认

10

mon_osd_max_op_age

描述

在发出 **HEALTH_WARN** 状态前，要完成的操作的最长时间（以秒为单位）。

类型

浮点值

默认

32.0

osd_pg_bits

描述

每个 Ceph OSD 守护进程的放置组位。

类型

32 位整数

默认

6

osd_pgp_bits

描述

用于放置目的 (PGP) 的每个 Ceph OSD 守护进程用于放置组的位数。

类型

32 位整数

默认

6

osd_crush_chooseleaf_type

描述

在 CRUSH 规则中，用于 **chooseleaf** 的 bucket 类型。使用等级排名，而不是名称。

类型

32 位整数

默认

***1**。通常，含有一个或多个 Ceph OSD 守护进程的主机。*

osd_pool_default_crush_replicated_ruleset

描述

创建复制池时要使用的默认 CRUSH 规则集。

类型

8 位整数

默认

0

osd_pool_erasure_code_stripe_unit

描述

*为纠删代码池设置对象条带的块的默认大小，以字节为单位。每个大小为 S 的对象将存储为 N 个条带，每个数据块都会有 **stripe unit** 个字节。**N * stripe unit** 字节的每个条带都将单独编码/解码。这个选项可以通过 profile 中的 **stripe_unit** 设置覆盖。*

类型

Unsigned 32 位整数

默认

4096

osd_pool_default_size

描述

*设置池中对象的副本数量。默认值与 **ceph osd pool set {pool-name} size {size}** 相同。*

类型

32 位整数

默认

3

osd_pool_default_min_size

描述

*为池中对象设置最少写入副本数，以确认对客户端的写操作。如果未达到最小值，Ceph 不会确认对客户端的写入。此设置确保以 **degraded** 模式运行时有最小副本数。*

类型

32 位整数

默认

***0**，表示没有特定最小值。如果为 **0**，最小为 **size - (size / 2)** 。*

osd_pool_default_pg_num

描述

*池的默认放置组数量。默认值为 **pg_num** 和 **mkpool**。*

类型

32 位整数

默认

32

osd_pool_default_pgp_num

描述

池放置的默认放置组数量。默认值为 ***pgp_num*** *和* ***mkpool****。PG 和 PGP 应该相等。*

类型

32 位整数

默认

0

osd_pool_default_flags

描述

新池的默认标记。

类型

32 位整数

默认

0

osd_max_pgls

描述

要列出的最大放置组数量。请求大量客户端可以连接 Ceph OSD 守护进程。

类型

unsigned 64 位整数

默认

1024

备注

默认应该是正常的。

osd_min_pg_log_entries

描述

修剪日志文件时要维护的最小放置组日志数量。

类型

32 位整数（Unsigned）

默认

250

osd_default_data_pool_replay_window

描述

OSD 等待客户端重播请求的时间（以秒为单位）。

类型

32 位整数

默认

45

附录 F. OBJECT STORAGE DAEMON (OSD) 配置选项

以下是部署期间可以设置的 Ceph Object Storage Daemon (OSD) 配置选项。

您可以使用 **ceph config set osd CONFIGURATION_OPTION VALUE** 命令设置这些配置选项。

osd_uuid

描述

Ceph OSD 的通用唯一识别符 (UUID) 。

类型

UUID

默认

UUID。

备注

osd uuid 应用到单个 Ceph OSD。 **fsid** 应用到整个集群。

osd_data

描述

OSD 数据路径。在部署 Ceph 时，您必须创建该目录。在此挂载点上挂载 OSD 数据的驱动器。

> IMPORTANT: Red Hat does not recommend changing the default.

类型

字符串

默认

/var/lib/ceph/osd/$cluster-$id

osd_max_write_size

描述

以 MB 为单位的最大写入大小。

类型

32 位整数

默认

90

osd_client_message_size_cap

描述

内存中允许的最大客户端数据消息。

类型

64 位 Unsigned 整数

默认

500MB 默认。**500*1024L*1024L**

osd_class_dir

描述

RADOS 类插件的类路径。

类型

字符串

默认

$libdir/rados-classes

osd_max_scrubs

描述

Ceph OSD 同步清理操作的最大数量。

类型

32 位整数

默认

1

osd_scrub_thread_timeout

描述

刮除线程超时前需要经过的最大时间（以秒为单位）。

类型

32 位整数

默认

60

osd_scrub_finalize_thread_timeout

描述

刮除完成线程超时前需要经过的最大时间（以秒为单位）。

类型

32 位整数

默认

60*10

osd_scrub_begin_hour

描述

这会将清理限制为一天或之后的这个小时。使用 **osd_scrub_begin_hour = 0** 和 **osd_scrub_end_hour = 0** 以允许清理整个一天。与 **osd_scrub_end_hour** 一起，它们定义了一个时间窗，在其中执行清理。但是，无论时间窗是否允许，只要放置组的清理间隔超过 **osd_scrub_max_interval**，就会执行清理。

类型

整数

默认

0

允许的范围

[0,23]

osd_scrub_end_hour

描述

这会将清理限制为比此问题更早的小时。使用 **osd_scrub_begin_hour = 0** 和 **osd_scrub_end_hour = 0** 允许在整个一天进行清理。与 **osd_scrub_begin_hour** 一起，它们定义了一个时间窗，在其中执行清理。但是，无论时间窗是否允许，只要放置组的清理间隔超过 **osd_scrub_max_interval**，就会执行清理。

类型

整数

默认

0

允许的范围

[0,23]

osd_scrub_load_threshold

描述

最大负载。当系统负载（由 **getloadavg ()** 功能定义）超过这个数值时，Ceph 不会进行刮除。默认为 **0.5**。

类型

浮点值

默认

0.5

osd_scrub_min_interval

描述

当 Red Hat Ceph Storage 集群负载较低时，清理 Ceph OSD 的最小间隔（以秒为单位）。

类型

浮点值

默认

每天一次。**60*60*24**

osd_scrub_max_interval

描述

清理 Ceph OSD 所需负载时的最长时间（以秒为单位）。

类型

浮点值

默认

每周一次。**7*60*60*24**

osd_scrub_interval_randomize_ratio

描述

使用这个比率，在 **osd scrub min interval** 和 **osd scrub max interval** 间随机化调度的刮除。

类型

浮点值

默认

0.5.

mon_warn_not_scrubbed

描述

osd_scrub_interval 后的秒数，以警告任何未清理的 PG。

类型

整数

默认

0（无警告）。

osd_scrub_chunk_min

描述

对象存储被分区为以哈希界限结尾的块。对于块清理，Ceph 每次清理对象的一个块，且会阻止对这个块的写入。**osd scrub chunk min** 设置表示要清理的最小块数量。

类型

32 位整数

默认

5

osd_scrub_chunk_max

描述

清理的最大块数量。

类型

32 位整数

默认

25

osd_scrub_sleep

描述

深度清理操作之间休眠的时间。

类型

浮点值

默认

0（或关闭）。

osd_scrub_during_recovery

描述

允许在恢复期间进行清理。

类型

Bool

默认

false

osd_scrub_invalid_stats

描述

强制执行额外的清理，以修复标记为无效统计数据。

类型

Bool

默认

true

osd_scrub_priority

描述

控制清理操作与客户端 I/O 的队列优先级。

类型

Unsigned 32 位整数

默认

5

osd_requested_scrub_priority

描述

在工作队列中为用户请求清理设置的优先级。如果这个值要小于 **osd_client_op_priority**，可以在清理阻止客户端操作时将其提高到 **osd_client_op_priority** 的值。

类型

Unsigned 32 位整数

默认

120

osd_scrub_cost

描述

以 MB 为单位清理操作的成本，用于队列调度目的。

类型

Unsigned 32 位整数

默认

52428800

osd_deep_scrub_interval

描述

深度清理的时间间隔，即完全读取所有数据。**osd scrub load threshold** 参数不会影响此设置。

类型

浮点值

默认

每周一次。**60*60*24*7**

osd_deep_scrub_stride

描述

在进行深度清理时读取大小。

类型

32 位整数

默认

512 KB.**524288**

mon_warn_not_deep_scrubbed

描述

osd_deep_scrub_interval 后的秒数，以警告任何未清理的 PG。

类型

整数

默认

0（无警告）

osd_deep_scrub_randomize_ratio

描述

清理的速率会随机变得深度清理（即使 **osd_deep_scrub_interval** 已通过）。

类型

浮点值

默认

0.15 或 15%

osd_deep_scrub_update_digest_min_age

描述

在清理更新整个对象摘要前，旧对象需要有多少秒。

类型

整数

默认

7200（120 小时）

osd_deep_scrub_large_omap_object_key_threshold

描述

当您遇到多个 OMAP 密钥的对象时，会发出警告。

类型

整数

默认

200000

osd_deep_scrub_large_omap_object_value_sum_threshold

描述

当您遇到多个 OMAP 键字节超过这个对象时，会发出警告。

类型

整数

默认

1 G

osd_delete_sleep

描述

下一次删除事务前休眠的时间（以秒为单位）。这会节流 PG 删除过程。

类型

浮点值

默认

0.0

osd_delete_sleep_hdd

描述

为 HDD 下一次删除事务前休眠的时间（以秒为单位）。

类型

浮点值

默认

5.0

osd_delete_sleep_ssd

描述

SSD 下一次删除事务前休眠的时间（以秒为单位）。

类型

浮点值

默认

1.0

osd_delete_sleep_hybrid

描述

当 Ceph OSD 数据位于 HDD 和 OSD 日志或 WAL 和 DB 为 SSD 上时，下一个删除事务前需要休眠的时间（以秒为单位）。

类型

浮点值

默认

1.0

osd_op_num_shards

描述

客户端操作的分片数量。

类型

32 位整数

默认

0

osd_op_num_threads_per_shard

描述

客户端操作的每个分片的线程数量。

类型

32 位整数

默认

0

osd_op_num_shards_hdd

描述

HDD 操作的分片数量。

类型

32 位整数

默认

5

osd_op_num_threads_per_shard_hdd

描述

每个分片用于 HDD 操作的线程数量。

类型

32 位整数

默认

1

osd_op_num_shards_ssd

描述

SSD 操作的分片数量。

类型

32 位整数

默认

8

osd_op_num_threads_per_shard_ssd

描述

用于 SSD 操作的每个分片的线程数量。

类型

32 位整数

默认

2

osd_op_queue

描述

设置在 Ceph OSD 中优先操作的队列类型。需要重启 OSD 守护进程。

类型

字符串

默认

wpq

有效选择

wpq,mclock_scheduler,debug_random

重要

mClock OSD 调度程序只是一个技术预览功能。红帽产品服务级别协议（SLA）不支持技术预览功能，且其功能可能并不完善，因此红帽不建议在生产环境中使用它们。这些技术预览功能可以使用户提早试用新的功能，并有机会在开发阶段提供反馈意见。如需了解更多详细信息，请参阅红帽技术预览功能的支持范围。

osd_op_queue_cut_off

描述

选择哪些优先级操作发送到严格的队列，哪些发送到正常的队列。需要重启 OSD 守护进程。low 设置将所有复制和更高操作发送到严格的队列，而 high 选项则仅将复制确认操作及更高的操作发送到严格的队列。

当集群中的某些 Ceph OSD 非常忙碌时，特别是与 **osd_op_queue** 设置中的 **wpq** 选项合并时，高设置将有所帮助。在处理复制流量非常忙碌的 Ceph OSD 可以在这些 OSD 上分离主客户端流量，而无需这些设置。

类型

字符串

默认

high

有效选择

low,high,debug_random

osd_client_op_priority

描述

为客户端操作设置的优先级。它相对于 **osd recovery op priority**。

类型

32 位整数

默认

63

有效范围

1-63

osd_recovery_op_priority

描述

恢复操作设置的优先级。它相对于 **osd client op priority**。

类型

32 位整数

默认

3

有效范围

1-63

osd_op_thread_timeout

描述

Ceph OSD 操作线程超时（以秒为单位）。

类型

32 位整数

默认

15

osd_op_complaint_time

描述

在经过指定秒数后，某个操作会变得令人满意。

类型

浮点值

默认

30

osd_disk_threads

描述

用于执行后台磁盘密集型 OSD 操作的磁盘线程数量，如清理和 snap 修剪。

类型

32 位整数

默认

1

osd_op_history_size

描述

要跟踪的最大完成操作数。

类型

32-bit Unsigned 整数

默认

20

osd_op_history_duration

描述

要跟踪的最旧的已完成操作。

类型

32-bit Unsigned 整数

默认

600

osd_op_log_threshold

描述

一次显示多少个操作日志。

类型

32 位整数

默认

5

osd_op_timeout

描述

运行 OSD 操作超时的时间（以秒为单位）。

类型

整数

默认

0

重要

*不要设置 **osd op timeout** 选项，除非您的客户端可以处理后果。例如，在虚拟机中运行的客户端设置此参数可能会导致数据崩溃，因为虚拟机将此超时解释为硬件故障。*

osd_max_backfills

描述

允许从一个 OSD 或单个 OSD 允许的最大回填操作数。

类型

64-bit Unsigned 整数

默认

1

osd_backfill_scan_min

描述

每次回填扫描的最小对象数量。

类型

32 位整数

默认

64

osd_backfill_scan_max

描述

每次回填扫描的最大对象数量。

类型

32 位整数

默认

512

osd_backfill_full_ratio

描述

当 Ceph OSD 的全满比率超过这个值时，拒绝接受回填请求。

类型

浮点值

默认

0.85

osd_backfill_retry_interval

描述

在重试回填请求前等待的秒数。

类型

双

默认

30.000000

osd_map_dedup

描述

启用删除 OSD map 中的重复项。

类型

布尔值

默认

true

osd_map_cache_size

描述

以 MB 为单位的 OSD map 缓存的大小。

类型

32 位整数

默认

50

osd_map_cache_bl_size

描述

OSD 守护进程中的内存中 OSD map 缓存的大小。

类型

32 位整数

默认

50

osd_map_cache_bl_inc_size

描述

内存中 OSD 映射缓存在 OSD 守护进程中递增的大小。

类型

32 位整数

默认

100

osd_map_message_max

描述

每个 MOSDMap 消息允许的最大映射条目。

类型

32 位整数

默认

40

osd_snap_trim_thread_timeout

描述

在超时 snap trim 线程前的最大时间（以秒为单位）。

类型

32 位整数

默认

60*60*1

osd_pg_max_concurrent_snap_trims

描述

并行 snap 修剪/PG 的最大数量。这将控制每个 PG 要一次修剪的对象数量。

类型

32 位整数

默认

2

osd_snap_trim_sleep

描述

在 PG 发布的每个修剪操作之间插入一个 sleep。

类型

32 位整数

默认

0

osd_snap_trim_sleep_hdd

描述

为 HDD 下一次快照修剪前休眠的时间（以秒为单位）。

类型

浮点值

默认

5.0

osd_snap_trim_sleep_ssd

描述

下一次快照修剪操作前休眠的时间，包括 NVMe。

类型

浮点值

默认

0.0

osd_snap_trim_sleep_hybrid

描述

当 OSD 数据位于 HDD 且 OSD 日志或 WAL 和 DB 位于 SSD 上时，下一个快照修剪操作前需要休眠的时间（以秒为单位）。

类型

浮点值

默认

2.0

osd_max_trimming_pgs

描述

修剪 PG 的最大数量

类型

32 位整数

默认

2

osd_backlog_thread_timeout

描述

backlog 线程超时前需要经过的最大时间（以秒为单位）。

类型

32 位整数

默认

60*60*1

osd_default_notify_timeout

描述

OSD 默认通知超时（以秒为单位）。

类型

32 位整数（Unsigned）

默认

30

osd_check_for_log_corruption

描述

检查日志文件是否存在损坏。计算的代价可能会比较高。

类型

布尔值

默认

false

osd_remove_thread_timeout

描述

在超时删除 OSD 线程前的最大时间（以秒为单位）。

类型

32 位整数

默认

60*60

osd_command_thread_timeout

描述

命令线程超时前需要经过的最大时间（以秒为单位）。

类型

32 位整数

默认

10*60

osd_command_max_records

描述

限制丢失对象的数量。

类型

32 位整数

默认

256

osd_auto_upgrade_tmap

描述

在旧对象为 **omap** 使用 **tmap**。

类型

布尔值

默认

true

osd_tmapput_sets_users_tmap

描述

仅使用 **tmap** 进行调试。

类型

布尔值

默认

false

osd_preserve_trimmed_log

描述

保留会修剪的日志文件，但会占用更多磁盘空间。

类型

布尔值

默认

false

osd_recovery_delay_start

描述

在对等点完成后，开始恢复对象前，Ceph 延迟指定的秒数。

类型

浮点值

默认

0

osd_recovery_max_active

描述

一次每个 OSD 活跃的恢复请求数。更多请求会加快恢复速度，但请求会增加集群中的负载。

类型

32 位整数

默认

0

osd_recovery_max_active_hdd

描述

如果主设备是 HDD，则一次每个 Ceph OSD 的活跃恢复请求数。

类型

整数

默认

3

osd_recovery_max_active_ssd

描述

如果主设备是 SSD，则一次每个 Ceph OSD 的活动恢复请求数。

类型

整数

默认

10

osd_recovery_sleep

描述

下一次恢复或回填操作前休眠的时间（以秒为单位）。增加这个值会减慢恢复操作速度，而客户端操作会受到影响。

类型

浮点值

默认

0.0

osd_recovery_sleep_hdd

描述

HDD 下一次恢复或回填操作前休眠的时间（以秒为单位）。

类型

浮点值

默认

0.1

osd_recovery_sleep_ssd

描述

SSD 下一次恢复或回填操作前休眠的时间（以秒为单位）。

类型

浮点值

默认

0.0

osd_recovery_sleep_hybrid

描述

当 Ceph OSD 数据位于 HDD，并且 OSD 日志或 WAL 和 DB 位于 SSD 上时，下一次恢复或回填操作前需要休眠的时间（以秒为单位）。

类型

浮点值

默认

0.025

osd_recovery_max_chunk

描述

要推送的数据恢复块的最大大小。

类型

64 位 Unsigned 整数

默认

8388608

osd_recovery_threads

描述

恢复数据的线程数量。

类型

32 位整数

默认

1

osd_recovery_thread_timeout

描述

超时恢复线程前的最大时间（以秒为单位）。

类型

32 位整数

默认

30

osd_recover_clone_overlap

描述

在恢复期间保留克隆重叠。应始终设为 **true**。

类型

布尔值

默认

true

rados_osd_op_timeout

描述

RADOS 在从 RADOS 操作返回错误之前等待来自 OSD 的响应的秒数。值为 0 表示没有限制。

类型

双

默认

0

附录 G. CEPH 监控器和 OSD 配置选项

修改 heartbeat 设置时，请将它们包含在 Ceph 配置文件的 **[global]** 部分中。

mon_osd_min_up_ratio

描述

Ceph 在标记 Ceph OSD 守护进程为 **down** 前的最小 **up** Ceph OSD 守护进程比率。

类型

双

默认

.3

mon_osd_min_in_ratio

描述

在 Ceph 将 Ceph OSD 守护进程标记为 **out** 之前， **in** Ceph OSD 守护进程的最小比率。

类型

双

默认

0.750000

mon_osd_laggy_halflife

描述

laggy 估算将会衰变的秒数。

类型

整数

默认

60*60

mon_osd_laggy_weight

描述

laggy 估算衰变中新样本的权重。

类型

双

默认

0.3

mon_osd_laggy_max_interval

描述

设置 **lagggy_interval** 的值（以秒为单位）。monitor 使用调整性方法来评估特定 OSD 的 **laggy_interval**。这个值将用于计算该 OSD 的宽限期。

类型

整数

默认

300

mon_osd_adjust_heartbeat_grace

描述

如果设置为 ***true***，*Ceph 将根据* ***laggy*** *估算进行扩展。*

类型

布尔值

默认

true

mon_osd_adjust_down_out_interval

描述

如果设置为 ***true***，*Ceph 将基于* ***laggy*** *估算进行扩展。*

类型

布尔值

默认

true

mon_osd_auto_mark_in

描述

Ceph 会将任何启动 Ceph OSD 守护进程标记为 ***in*** *Ceph Storage 集群。*

类型

布尔值

默认

false

mon_osd_auto_mark_auto_out_in

描述

Ceph 将自动标记为 ***out*** *Ceph Storage Cluster 的引导 Ceph OSD 守护进程标记为* ***in*** *集群。*

类型

布尔值

默认

true

mon_osd_auto_mark_new_in

描述

Ceph 会将引导新的 Ceph OSD 守护进程标记为 ***in*** *Ceph Storage 集群。*

类型

布尔值

默认

true

mon_osd_down_out_interval

描述

当一个 Ceph OSD 守护进程没有响应时，Ceph 在将其标记为 ***down*** *和* ***out*** *前等待的时间。*

类型

32 位整数

默认

600

mon_osd_downout_subtree_limit

描述

Ceph 将自动标记为 ***out*** *的最大 CRUSH 单元类型。*

类型

字符串

默认

rack

mon_osd_reporter_subtree_level

描述

此设置为报告 OSD 定义父 CRUSH 单元类型。如果 OSD 找到没有响应的对等点，OSD 会向监控器发送失败报告。monitor 标记报告的 OSD ***down****，并在宽限期后设置为* ***out****。*

类型

字符串

默认

主机

mon_osd_report_timeout

描述

在将没有响应的 Ceph OSD Daemons 声明为 ***down*** *前经过的宽限期（以秒为单位）。*

类型

32 位整数

默认

900

mon_osd_min_down_reporters

描述

报告 ***down*** *Ceph OSD 守护进程需要的最小 Ceph OSD 守护进程的数量。*

类型

32 位整数

默认

2

osd_heartbeat_address

描述

用于心跳的 Ceph OSD 守护进程的网络地址。

类型

地址

默认

主机地址。

osd_heartbeat_interval

描述

Ceph OSD 守护进程如何 ping 对等点（以秒为单位）。

类型

32 位整数

默认

6

osd_heartbeat_grace

描述

当 Ceph OSD 守护进程未显示心跳后需要经过的时间 Ceph Storage Cluster 才认为它的状态为 **down**。

类型

32 位整数

默认

20

osd_mon_heartbeat_interval

描述

如果没有 Ceph OSD 守护进程同级服务器，Ceph OSD 守护进程会如何 ping Ceph 监控器。

类型

32 位整数

默认

30

osd_mon_report_interval_max

描述

Ceph OSD 守护进程在报告到 Ceph 监控器之前可以等待的时间（以秒为单位）。

类型

32 位整数

默认

120

osd_mon_report_interval_min

描述

在向 Ceph monitor 报告前，Ceph OSD 守护进程可以从启动或其他可报告事件等待的最少秒数。

类型

32 位整数

默认

5

有效范围

应小于 ***osd mon report interval max***

osd_mon_ack_timeout

描述

等待 Ceph Monitor 的秒数，以确认对统计数据的请求。

类型

32 位整数

默认

30

附录 H. CEPH 刮除选项

Ceph 通过清理放置组来确保数据完整性。以下是您可以调整的 Ceph 清理选项，以增加或减少刮除操作。

您可以使用 **ceph config set global CONFIGURATION_OPTION VALUE** 命令设置这些配置选项。

mds_max_scrub_ops_in_progress

描述

并行执行的最大清理操作数量。您可以使用 **ceph config set mds_max_scrub_ops_in_progress VALUE** 命令设置这个值。

类型

整数

默认

5

osd_max_scrubs

描述

Ceph OSD 守护进程的同步清理操作的最大数量。

类型

整数

默认

1

osd_scrub_begin_hour

描述

清理开始的特定小时。和 **osd_scrub_end_hour** 一起，您可以定义一个可发生清理的时间窗。使用 **osd_scrub_begin_hour = 0** 和 **osd_scrub_end_hour = 0** 以允许清理整个一天。

类型

整数

默认

0

允许的范围

[0, 23]

osd_scrub_end_hour

描述

清理结束的特定小时。与 **osd_scrub_begin_hour** 一起，您可以定义一个时间窗，在其中执行清理。使用 **osd_scrub_begin_hour = 0** 和 **osd_scrub_end_hour = 0** 允许在整个一天进行清理。

类型

整数

默认

0

允许的范围

[0, 23]

osd_scrub_begin_week_day

描述

清理开始的具体日期。0 = Sunday, 1 = Monday 等.除了 "osd_scrub_end_week_day" 外，您还可以定义一个可以发生清理的时间窗。使用 **osd_scrub_begin_week_day = 0** 和 **osd_scrub_end_week_day = 0** 以允许对整个星期的清理。

类型

整数

默认

0

允许的范围

[0, 6]

osd_scrub_end_week_day

描述

这将定义清理结束的日期。0 = Sunday, 1 = Monday 等.与 **osd_scrub_begin_week_day** 一起，它们定义了一个时间窗，其中执行清理。使用 **osd_scrub_begin_week_day = 0** 和 **osd_scrub_end_week_day = 0** 以允许对整个星期的清理。

类型

整数

默认

0

允许的范围

[0, 6]

osd_scrub_during_recovery

描述

在恢复期间允许清理。把它设置为 **false** 可禁用调度新的清理和深度清理，同时存在活跃的恢复。已在运行的刮除会继续，对降低忙碌存储集群上的负载很有用。

类型

布尔值

默认

false

osd_scrub_load_threshold

描述

规范化最大负载。当系统负载（由 getloadavg () / 在线 CPU 的数量指定）的数量高于这一定义的数量时，刮除不会发生。

类型

浮点值

默认

0.5

osd_scrub_min_interval

描述

当 Ceph 存储集群负载较低时，清理 Ceph OSD 守护进程的最小间隔（以秒为单位）。

类型

浮点值

默认

1 天

osd_scrub_max_interval

描述

清理 Ceph OSD 守护进程时的最大间隔，以秒为单位。

类型

浮点值

默认

7 天

osd_scrub_chunk_min

描述

在单个操作期间，要刮除的对象存储块的最小数量。Ceph 块在清理期间写入单个块。

type

整数

默认

5

osd_scrub_chunk_max

描述

在单个操作期间，要刮除的对象存储块的最大数量。

type

整数

默认

25

osd_scrub_sleep

描述

在刮除下一块组前需要经过的睡眠时间。增加这个值会减慢刮除的整体速率，因此客户端操作受到的影响较低。

type

浮点值

默认

0.0

osd_scrub_extended_sleep

描述

在清理超时或秒内注入延迟的时间。

type

浮点值

默认

0.0

osd_scrub_backoff_ratio

描述

调度清理的 backoff 比率。这是不调度清理的百分比，66% 表示 1 不再使用 3 ticks 调度清理。

type

浮点值

默认

0.66

osd_deep_scrub_interval

描述

深度清理的间隔，请完全读取所有数据。**osd_scrub_load_threshold** 不会影响此设置。

type

浮点值

默认

7 天

osd_debug_deep_scrub_sleep

描述

在深度清理 IO 期间注入昂贵的睡眠状态，以便更轻松地降低抢占。

type

浮点值

默认

0

osd_scrub_interval_randomize_ratio

描述

在为放置组调度下一个清理作业时，添加一个随机延迟到 **osd_scrub_min_interval**。延迟是一个随机值，小于 **osd_scrub_min_interval** * **osd_scrub_interval_randomized_ratio**。默认设置会在允许的时间窗口 **[1, 1.5]** * **osd_scrub_min_interval** 中分配清理。

type

浮点值

默认

0.5

osd_deep_scrub_stride

描述

在进行深度清理时读取大小。

type

size

默认

512 KB

osd_scrub_auto_repair_num_errors

描述

如果发现多个错误，则不会进行自动修复。

type

整数

默认

5

osd_scrub_auto_repair

描述

当清理或深度清理错误时，将其设置为 **true** 可启用自动放置组 (PG) 修复。但是，如果发现的错误数量超过了 **osd_scrub_auto_repair_errors**，则不允许进行修复。

type

布尔值

默认

false

osd_scrub_max_preemptions

描述

在阻止客户端 IO 完成清理前，设置因为客户端操作而需要抢占深度清理的最大次数。

type

整数

默认

5

osd_deep_scrub_keys

描述

在深度清理期间从对象读取的密钥数量。

type

整数

默认

1024

附录 I. BLUESTORE 配置选项

如下为 Ceph BlueStore 配置选项，可以在部署期间配置。

注意

此列表没有完成。

rocksdb_cache_size

描述

以 MB 为单位的 RocksDB 缓存的大小。

类型

32 位整数

默认

512

bluestore_throttle_bytes

描述

用户节流输入或输出(I/O)提交前的最大字节数。

类型

大小

默认

64 MB

bluestore_throttle_deferred_bytes

描述

用户节流 I/O 提交前延迟写入的最大字节数。

类型

大小

默认

128 MB

bluestore_throttle_cost_per_io

描述

每个 I/O 增加的开销（以字节为单位）。

类型

大小

默认

0 B

bluestore_throttle_cost_per_io_hdd

描述

HDD 的默认 **bluestore_throttle_cost_per_io** 值。

类型

未签名的整数

默认

67 000

bluestore_throttle_cost_per_io_ssd

描述

SSD 的默认 ***bluestore_throttle_cost_per_io*** *值。*

类型

未签名的整数

默认

4 000

bluestore_debug_enforce_settings

描述

HDD *强制实施用于轮转驱动器之上的* BlueStore *设置。* ***SSD*** *强制执行用于固态硬盘之上的设置*

类型

默认,hdd,ssd

默认

default

注意

在更改 ***bluestore_debug_enforce_settings*** *选项后，重启 OSD。*